Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Spatiotemporal modeling of cyanobacterial cell counts in the Billings Reservoir using Sentinel-2 and machine learning

Rafael Ribeiro de Lima

rafaelrlima@usp.br

University of São Paulo

Walter Jaimes Salcedo

University of São Paulo

Research Article

Keywords: cyanobacteria, cell count, remote sensing, Sentinel-2, machine learning, Billings Reservoir

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11815730/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Spatiotemporal modeling of cyanobacterial cell counts in the Billings Reservoir using Sentinel-2 and machine learning

Rafael Ribeiro de Lima[1] · Walter Jaimes Salcedo[1]

[1] Polytechnic School, University of São Paulo (USP), São Paulo, SP, Brazil; rafaelrlima@usp.br; walter.salcedo@usp.br.
Corresponding author: Rafael Ribeiro de Lima

Abstract

The integration of satellite observations and field monitoring provides opportunities to investigate cyanobacteria in inland waters. This study examined the number of cyanobacterial cells (NCC) in the Billings Reservoir, located in the state of São Paulo, Brazil, using CETESB monitoring records, Sentinel-2 Level-2A imagery, spectral indices, and machine learning. Data from five monitoring stations were matched to satellite acquisitions within seven days; reflectance was summarized using 3 × 3-pixel windows. After spectral and temporal feature extraction, the station series were regularized to monthly intervals, yielding 515 station-month records, including 403 months without observed NCC before multivariate reconstruction with a diffusion-inspired LSTM architecture. Fourteen regression configurations were compared using 455 processed records from 2017–2024 for development and 60 processed records from 2025 for a chronological comparison. The Keras-based MLP obtained the highest R^2 (0.8525), with RMSE of 222,998,226.43 and MAE of 72,310,835.10 cells/mL relative to the processed monthly reference. The model was also applied exploratorily to a December 2025 Sentinel-2 scene to visualize the spatial distribution of estimated NCC classes. Because the monthly reference contains reconstructed values and reconstruction preceded the chronological split, these metrics do not constitute independent validation against laboratory-only observations. Likewise, the spatial outputs are model estimates, not direct measurements of cyanobacterial cells or cyanotoxins.

Keywords: cyanobacteria; cell count; remote sensing; Sentinel-2; machine learning; Billings Reservoir.

1 Introduction

Cyanobacterial blooms are of ecological and public health importance in inland waters. Eutrophication, associated with nutrient availability and favorable environmental conditions, can increase bloom frequency and intensity; some blooms involve organisms capable of producing cyanotoxins (Codd 2000; O'Neil et al. 2012). In reservoirs used for water supply, monitoring cyanobacterial abundance is important for guiding water quality surveillance and prioritizing management actions.

Field sampling and laboratory analyses provide the reference observations required to quantify cyanobacterial cells, but they depend on discrete campaigns and available resources. Remote sensing complements these measurements by providing repeated, spatially distributed observations of surface optical properties. Interpretation remains challenging because aquatic reflectance depends on water, phytoplankton pigments, suspended particles, dissolved organic matter, and atmospheric and surface effects (Matthews 2011; Giardino et al. 2007). Accordingly, spectral indices are indirect proxies and should not be treated as cyanobacteria-specific measurements.

The Sentinel-2 Multispectral Instrument (MSI) acquires spectral bands in the visible, near-infrared (NIR), red-edge, and shortwave-infrared (SWIR) regions, at spatial resolutions of 10, 20, and 60 m (Drusch et al. 2012; European Space Agency 2021). Table 1 lists the MSI bands, their central wavelengths, and corresponding spatial resolutions. Mathematical combinations of these bands yield spectral indices, defined as ratios, differences, or transformations of reflectance at different wavelengths intended to enhance specific optical characteristics of observed targets. In aquatic monitoring, these indices are widely used as indirect indicators of water quality properties, including chlorophyll-a concentration, turbidity, algal biomass, and cyanobacterial blooms (Shi et al. 2019; Llodrà-Llabrés et al. 2023). Besides summarizing relevant spectral information, some formulations—particularly normalized ratios and indices based on spectral baselines—can reduce sensitivity to changes in illumination, viewing geometry, and certain atmospheric interferences, thereby improving estimate stability under specific conditions. However, these methods neither eliminate the need for atmospheric correction nor ensure the removal of image noise (Mishra and Mishra 2012; Hu 2009; Llodrà-Llabrés et al. 2023).

Table 1. Spectral characteristics of the MSI bands aboard Sentinel-2, including central wavelength and spatial resolution. Sources: Drusch et al. (2012) and European Space Agency (2021).

Characteristic	B01	B02	B03	B04	B05	B06	B07	B08	B8A	B09	B10	B11	B12
Central wavelength (nm)	443	490	560	665	705	740	783	842	865	945	1,375	1,610	2,190
Spatial resolution (m)	60	10	10	10	20	20	20	10	20	60	60	20	20

Among the indices used to monitor algal blooms, the Normalized Difference Chlorophyll Index (NDCI) exploits the contrast between red and red-edge bands associated with the spectral response of chlorophyll-a in productive and turbid waters (Mishra and Mishra 2012). The Maximum Chlorophyll Index (MCI) identifies spectral enhancement near 705 nm, associated with photosynthetic pigments and high phytoplankton biomass (Binding et al. 2013). The Floating Algae Index (FAI) uses a spectral baseline between the red and SWIR bands to highlight increased NIR reflectance associated with surface algal material, showing lower sensitivity to certain environmental and observation-related variations than conventional vegetation indices (Hu 2009). The Surface Algal Bloom Index (SABI), in turn, was developed to assist in identifying surface algal accumulations from spectral band combinations (Alawadi 2010). Combining these features captures different optical characteristics of water and provides potentially informative predictors for machine learning models. However, these spectral responses are not exclusive to cyanobacteria and do not directly equate to laboratory determinations of cell counts, making their association with in situ observations necessary for developing and evaluating predictive models.

Machine learning methods can combine multiple predictors to represent nonlinear relationships between reflectance and water quality variables (Hastie et al. 2009; Reichstein et al. 2019). In Brazil, Pompêo et al. (2021) investigated Sentinel-2-based estimates of cyanobacterial cell counts in the Cantareira System; therefore, estimating this variable from Sentinel-2 is not unprecedented. CyFi likewise combines satellite imagery and machine learning to estimate cyanobacterial density in inland waters (Dorne et al. 2024). The present study contributes a reservoir-specific application to Billings, integrating spectral and temporal feature engineering, monthly reconstruction, comparative regression modeling, and exploratory spatial visualization. Its quantitative evaluation concerns a processed reference series containing reconstructed NCC values.

This study aimed to develop and describe an integrated workflow combining Sentinel-2 imagery with CETESB records for spatiotemporal analysis of NCC in the Billings Reservoir. Specifically, it examined the use of spectral indices and temporal descriptors, the reconstruction of monthly series with missing observations, the comparative performance of different regression configurations on the processed reference, and the exploratory spatial application of the selected model.

2 Materials and methods

2.1 Study area and monitoring data

The study area comprises the Billings Reservoir, an urban tropical reservoir located in the São Paulo Metropolitan Region, Brazil, within the Alto Tietê Water Resources Management Unit (UGRHI 6). Five monitoring stations operated by the São Paulo State Environmental Agency (CETESB) were selected across different reservoir sectors: BILL02030, in the central water body approximately 1.5 km from the Pedreira Dam; BILL02100, toward the Bororé arm; BILL02500, beneath the Imigrantes Highway bridge; BILL02900, near the Billings–Pedras regulating dam (Summit Control); and BIRP00500, in the Rio Pequeno arm approximately 2 km upstream of the Caminhos do Mar Highway (CETESB n.d.).

Reference data were obtained from publicly available CETESB water quality monitoring records covering 2017–2025. The documentary sources were the Billings System Raw Water Quality Bulletins, which report analytical results for parameters monitored at each sampling station. These bulletins identify the water system, station code and location, sampling date, and water quality parameter values. Available parameters include the Number of Cyanobacterial Cells (NCC), expressed as cells per milliliter (cells/mL), along with chlorophyll-a, turbidity, dissolved oxygen, and total phosphorus. CETESB states that the analytical results are stored in its institutional InterÁguas database and made publicly accessible through the INFOÁGUAS system (CETESB n.d.).

To construct the dataset, NCC records from the five selected stations were retained together with the station identifier, sampling date, and CETESB-reported value. Records were ordered chronologically and linked to their geographic coordinates for subsequent matching with Sentinel-2 multispectral imagery. The coordinates, in decimal degrees of latitude and longitude, were BILL02030 (−23.717778; −46.664167), BILL02100 (−23.754444; −46.644444), BILL02500 (−23.790833; −46.598333), BILL02900 (−23.821111; −46.525000), and BIRP00500 (−23.791111; −46.470556). Figure 1 shows the spatial distribution of these stations.

The published results correspond to actual sampling dates and do not originally constitute continuous monthly time series. Some months have no records because of sampling frequency or unavailable results. Such gaps were distinguished from observed values and were not interpreted as zero counts. The records were subsequently integrated with eligible spectral data and organized into monthly time series for each monitoring station, following the temporal matching and missing-data

reconstruction procedures described below. Thus, CETESB-reported NCC values served as the original observational reference, whereas subsequently reconstructed values were treated as estimates produced by computational processing.

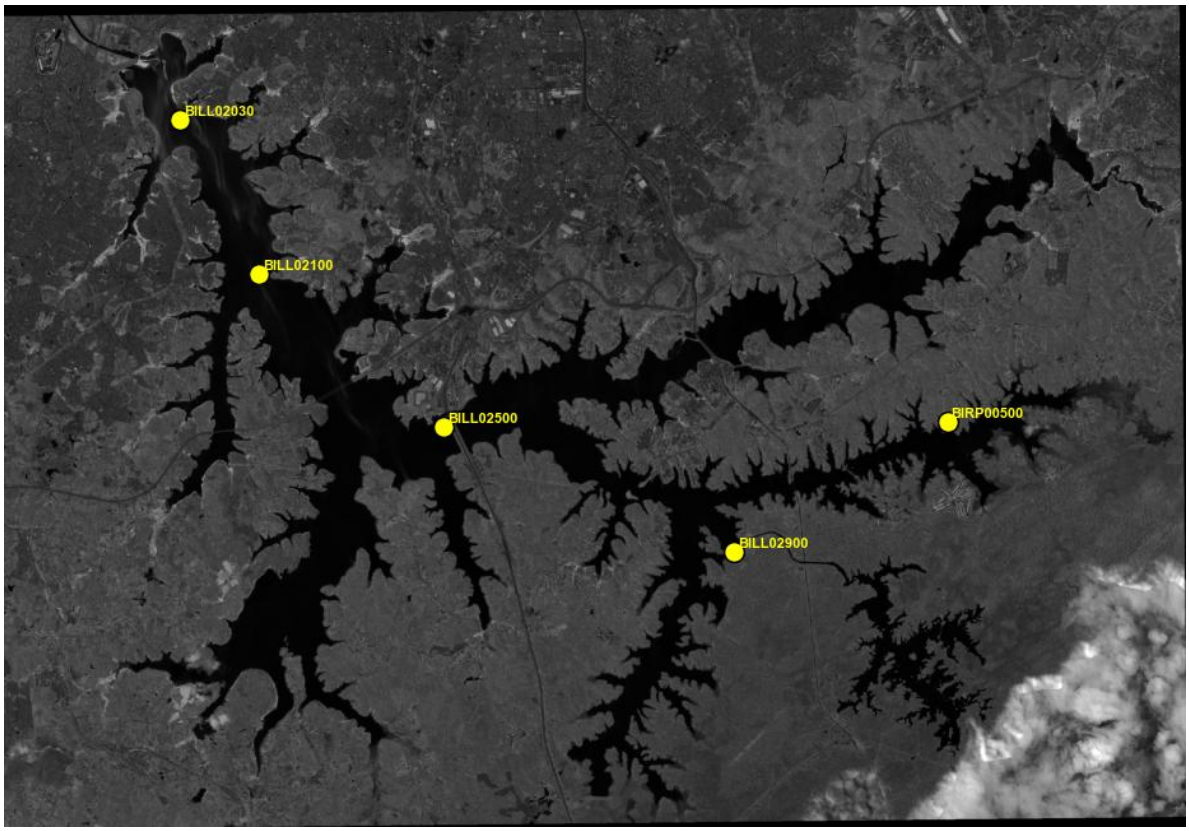


Figure 1. Billings Reservoir and locations of the five CETESB monitoring stations used to match observed NCC with Sentinel-2 imagery: BILL02030, BILL02100, BILL02500, BILL02900, and BIRP00500.

The workflow summarized in Figure 2 comprised six stages: (i) acquisition of CETESB NCC monitoring records and Sentinel-2 Level-2A imagery; (ii) image preprocessing, water-surface identification, cloud-related screening, and 3 × 3-pixel spectral extraction; (iii) calculation of spectral indices, integration with monitoring data, and organization of time series; (iv) multivariate reconstruction of missing values through a diffusion-inspired approach and complementary iterative imputation; (v) feature selection and regression-model comparison; and (vi) a chronological comparison on the processed 2025 series, followed by exploratory spatial application. The procedures are detailed below.

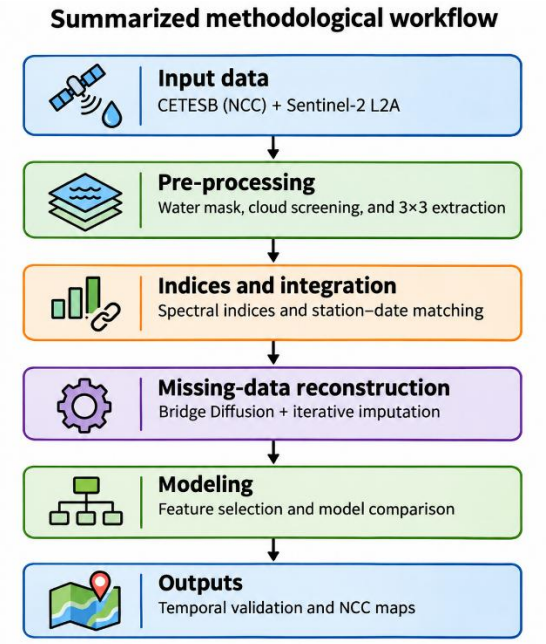


Figure 2. Summary of the methodological workflow for spatiotemporal modeling of the Number of Cyanobacterial Cells (NCC) in the Billings Reservoir, integrating CETESB data, Sentinel-2 imagery, time-series reconstruction, and machine learning. Source: prepared by the authors.

2.2 Image acquisition and quality control

Sentinel-2 Level-2A (L2A) products, representing Bottom-of-Atmosphere (BOA) surface reflectance, were used with the available spectral bands and the Scene Classification Layer (SCL). Level-2A processing contributes to atmospheric correction and surface-class identification, although it does not eliminate every source of uncertainty in inland waters. Scenes were organized by acquisition date, taking radiometric quality and potential matching with monitoring stations into account (Drusch et al. 2012; European Space Agency 2021).

The TIFF files were read for bands B01, B02, B03, B04, B05, B06, B07, B08, B8A, B09, B11, and B12, and for SCL, as available for each scene, together with the raster generated by the Ulyssys Water Quality Viewer (UWQV). NDWI and MNDWI were calculated from bands B03, B08, and B11. The water mask combined NDWI > 0.10 or MNDWI > 0.10 and, when available, the SCL water class (SCL = 6); SCL complemented the spectral classification. Cloud and shadow indicators were also produced as auxiliary processing information.

2.3 Temporal matching and spatial extraction

Station coordinates were reprojected to the raster spatial reference system when necessary. For each image, mean reflectance values were extracted from 3 × 3-pixel windows centered on the monitoring locations. This local aggregation reduces sensitivity to positional offsets and individual-pixel variability, but it also represents a spatial scale different from the pixel-level inference subsequently used.

Samples and images were matched by station code and calendar year–month, calculating the absolute difference between the acquisition and sampling dates. Matches were retained when NCC was available, the date difference was no more than seven days, and water classification was positive. When more than one record occurred for the same station-month, numerical values were averaged during monthly aggregation. This procedure combines field observations and satellite acquisitions close in time without assuming strict simultaneity.

2.4 Spectral indices and feature engineering

Feature engineering preceded monthly regularization and gap reconstruction. We calculated water-related indicators (NDWI, MNDWI, NDMI, AWEI, and WRI), pigment and algal biomass indices (NDCI, FAI, MCI, CIrededge, CIgreen, and SABI), ratios associated with turbidity and colored dissolved organic matter, and auxiliary brightness and cloud parameters. In the expressions, Bk denotes the reflectance of the corresponding Sentinel-2 band and $\varepsilon = 10^{-6}$ prevents division by zero. Table 2 presents the principal equations and their interpretive roles.

Table 2. Spectral indices and derived features used to characterize the water surface and water quality.

Index and interpretive use	Mathematical formulation
NDWI Water identification/characterization (McFeeters 1996)	$\frac{(B3 - B8)}{(B3 + B8)}$
MNDWI Differentiation of water surfaces (Xu 2006)	$\frac{(B3 - B11)}{(B3 + B11)}$
NDMI Moisture-related spectral response	$\frac{(B8 - B11)}{(B8 + B11)}$
NDTI Turbidity-related spectral indicator	$\frac{(B4 - B3)}{(B4 + B3)}$
NDCI Spectral proxy for chlorophyll-a/phytoplankton biomass (Mishra and Mishra 2012)	$\frac{(B5 - B4)}{(B5 + B4)}$
CI_green Feature associated with phytoplankton biomass	$\frac{B8}{B3} - 1$
AWEI_nsh Water characterization under non-shadowed conditions (Feyisa et al. 2014)	$4(\mathrm{B3} - \mathrm{B11}) - (0.25\mathrm{B8} + 2.75\mathrm{B12})$
AWEI_sh Water characterization under shadowed conditions (Feyisa et al. 2014)	$\mathrm{B2} + 2.5\mathrm{B3} - 1.5(\mathrm{B8} + \mathrm{B11}) - 0.25\mathrm{B12}$
FAI NIR departure from the spectral baseline; possible floating material signal (Hu 2009)	$\mathrm{FAI} = \mathrm{B8} - [\mathrm{B4} + (\mathrm{B11} - \mathrm{B4})\frac{(842 - 665)}{(1610 - 665)}]$
CIrededge Red-edge spectral index associated with pigments	$\frac{\mathrm{B8}}{(\mathrm{B5} + \varepsilon) - 1}$
MCI Spectral departure of band B5 from the baseline	$\mathrm{B5} - \{\mathrm{B4} + (\mathrm{B6} - \mathrm{B4}) \cdot \frac{(705 - 665)}{(740 - 665)}\}$
SABI Proxy for surface algal material	$\frac{(\mathrm{B8} - \mathrm{B4})}{(\mathrm{B2} + \mathrm{B3} + \varepsilon)}$
TURB_R_G Red/green ratio for turbidity	$\frac{\mathrm{B4}}{(\mathrm{B3} + \varepsilon)}$
CIRRUS_INDEX Conditional index; band B10 is not included in standard L2A products	$\frac{(\mathrm{B10} - \mathrm{B8})}{(\mathrm{B10} + \mathrm{B8} + \varepsilon)}$, only if B10 is available

The spectral indices and their mathematical formulations are summarized in Table 2. The calculated indices included NDCI, CIrededge, MCI, SABI, FAI, NDTI, CIgreen, and TURB_R_G, which characterize different optical properties of the water surface. The red and red-edge bands capture spectral contrasts associated with chlorophyll-a and elevated phytoplankton biomass, depending on the optical characteristics of the water (Gitelson et al. 2009; Mishra and Mishra 2012; Binding et al. 2013). The Floating Algae Index (FAI) uses a red-to-SWIR spectral baseline to highlight near-infrared reflectance associated with floating algal material (Hu 2009), while the Surface Algal Bloom Index (SABI) combines visible and near-infrared bands to support the detection of surface algal accumulations (Alawadi 2010).

The ulyssys feature was obtained by extracting the first band of a TIFF file exported from the Ulyssys Water Quality Viewer (UWQV), which provides Sentinel-2-derived representations of water optical properties associated with chlorophyll and suspended sediments (Zlinszky and Padányi-Gulyás 2020; Sentinel Hub n.d.). This feature was treated as a numerical predictor rather than an independent laboratory measurement of cyanobacterial cell counts (NCC). The CIRRUS_INDEX was conditionally calculated using bands B10 and B8; when B10 was unavailable in Sentinel-2 Level-2A products, the resulting missing values were handled through the imputation procedures described in the following section.

To incorporate temporal variability into the modeling process, four complementary transformations of the spectral indices were derived separately for each monitoring station using the available chronologically ordered observations: (i) anomaly, defined as the difference between the spectral index value at a given observation and the median of the corresponding station-specific series, allowing deviations from its central tendency to be characterized; (ii) z-score, expressing the deviation of an observed value from the series mean, normalized by its standard deviation; (iii) temporal difference, quantifying the change between consecutive available observations on the original scale of the index; and (iv) relative temporal change, representing the proportional variation between consecutive observations relative to the preceding value. Relative changes were expressed as proportions rather than percentages. These transformations were incorporated as derived features to represent spectral characteristics and their temporal variability. In this context, the term anomaly refers exclusively to a statistical deviation from the station-specific series median and does not, by itself, provide evidence of an anomalous environmental event.

These temporal features were incorporated into the multivariate dataset used for missing-data reconstruction and subsequent feature selection. Their inclusion aimed to provide the machine learning models with information about both spectral conditions and temporal dynamics potentially associated with variations in cyanobacterial abundance. Five of the eight predictors ultimately selected for model training were temporally derived features, highlighting the relevance of temporal information within the feature selection procedure.

2.5 Missing-data reconstruction

The extracted dataset was aggregated by station-month and expanded to continuous monthly intervals between the first and last month of each series. This yielded 515 records (four series of 102 months and one of 107), with 403 missing NCC values after regularization, representing 78.3% of monthly rows. The reconstruction matrix contained 78 eligible numerical variables, including bands, indices, temporal derivatives, and NCC; spatial identifiers and dates were retained as metadata. Relative position within NCC gaps was represented by an auxiliary gap_position variable.

Multivariate imputation, termed Bridge Diffusion in the implementation, used mean-and-standard-deviation scaling of numerical variables, a three-dimensional [station, month, feature] structure, and availability masks. The architecture comprised two LSTM layers with 128 units and a dropout rate of 0.20, followed by a temporal output layer. Training used the Adam optimizer and masked mean squared error over observed elements. Fifty perturbation steps followed the cosine schedule below.

$$\sigma_t = \cos[(\pi/2) \cdot t/50]$$

Training comprised 20 epochs distributed across two stages of 10 epochs each. The procedure is an iterative, diffusion-inspired reconstruction distinct from the probabilistic CSDI formulation presented by Tashiro et al. (2021).

During reconstruction, perturbation steps were traversed in reverse order, updating missing values while preserving originally observed entries. The output was the average of 20 reconstruction passes, without explicit additional random sampling in each pass. After inverse scaling, negative NCC estimates were clipped to zero; sequential filling was used as a residual fallback. The matrix contained both spectral bands and derived features, allowing spectral and temporal dependencies to be explored without assuming that only indices were imputed.

Expressing spectral relationships as indices before reconstruction incorporated physically motivated descriptors into the numerical matrix. FAI, NDCI, and MCI have established uses in characterizing pigments and algal material in productive waters (Hu 2009; Mishra and Mishra 2012; Binding et al. 2013). Some band ratios can attenuate shared multiplicative effects, whereas others can amplify noise when denominators are small. In this study, bands, indices, and temporal derivatives were included jointly; no controlled comparison was performed to determine whether this combination improves reconstruction relative to bands alone.

2.6 Feature selection

Supervised feature selection was performed on the pre-2025 development subset after construction of the processed monthly matrix. The routine considered 76 candidate variables and applied low-variance filtering, absolute Pearson correlation with NCC (minimum threshold of 0.10), iterative collinearity reduction using the variance inflation factor (VIF $\leq$ 10), and

permutation importance estimated with Random Forest. The procedure selected variables associated with NCC in the processed series, whose reference includes multivariate reconstruction (O'Brien 2007; Fisher et al. 2019).

The Random Forest model used for feature importance estimation was configured with 500 trees, a random seed of 42, and eight permutation repeats per variable. To rank the candidate features, a composite score was calculated by combining normalized absolute Pearson correlation with NCC (40%) and Random Forest permutation importance (60%). This approach was adopted to account for both the linear association with the target variable and the predictive contribution of each feature within a nonlinear model.

The final feature selection score was defined as:

$$Sj = 0.40\ rj^* + 0.60\ Ij^*$$

Here, S_j is the composite score for feature j, while rj* and Ij* denote normalized absolute Pearson correlation and normalized permutation importance, respectively.
Following the filtering procedures, 11 candidate features remained eligible, and the eight highest-ranked features were selected as inputs for the machine learning models. Table 3 presents their absolute correlations with NCC, Variance Inflation Factor (VIF) values, and permutation importances.

Table 3. Absolute correlation with NCC, VIF, and permutation importance for the eight selected features.

Selected feature	|r| with NCC	VIF	Permutation importance
FAI_pct	0.759445	7.100	0.091386
ulyssys	0.700491	1.919	0.051916
SABI_z	0.473664	9.736	0.040105
CIrededge_pct	0.620013	4.697	0.000000
MCI	0.403065	2.815	0.000000
CIRRUS_INDEX	0.143491	6.472	0.022637
TURB_R_G_z	0.245047	2.883	0.000000
NDCI_delta	0.239637	6.074	0.000000

The selected predictors were FAI_pct, ulyssys, SABI_z, CIrededge_pct, MCI, CIRRUS_INDEX, TURB_R_G_z, and NDCI_delta. They represent a combination of surface-biomass-sensitive spectral responses, red/red-edge information, turbidity-related variation, UWQV-derived data, and temporal descriptors. Zero permutation importance for some predictors means that their shuffling did not produce a measurable performance decrease in the specified procedure; their inclusion therefore reflects the composite ranking rather than independently established predictive value.

2.7 Supervised models and training

Twelve regression algorithms were evaluated—linear regression, Elastic Net, Partial Least Squares (PLS), Random Forest, Gradient Boosting, Extra Trees, K-Nearest Neighbors, Support Vector Regression (SVR), Gaussian Process Regression (GPR), an scikit-learn MLP, a Keras MLP, and an LSTM—along with two additional tree-based configurations using randomized hyperparameter search (RF_CV and GB_CV), for a total of 14 alternatives. These included linear models, distance-based methods, tree ensembles, and neural networks (Breiman 2001; Friedman 2001; Geurts et al. 2006; Hastie et al. 2009).

The MLP_Keras model consisted of three fully connected hidden layers with 128, 64, and 32 neurons, respectively, using the Rectified Linear Unit (ReLU) activation function. ReLU was employed to introduce nonlinearity into the network, enabling it to learn complex relationships between spectral and temporal predictors and cyanobacterial cell counts (NCC) that cannot be adequately represented by linear transformations alone. By retaining positive activation values while setting negative activation values to zero, ReLU also supports efficient gradient-based optimization and helps mitigate the vanishing gradient problem compared with saturating activation functions.

The network included a single-neuron linear output layer to generate continuous NCC estimates. Training was performed using the Adam optimizer with a learning rate of 0.001, median imputation of missing predictor values, standardization of input features and the target variable, a maximum of 200 epochs, and a batch size of 32 observations. Early stopping was implemented with a patience of 15 epochs, restoring the model weights associated with the lowest validation loss to reduce the risk of overfitting. Predictions were subsequently transformed back to the original NCC scale. Because the output layer did not impose a nonnegativity constraint, negative predictions could occur, although such values have no physical interpretation as cell counts.

The configuration designated LSTM received the input features in a three-dimensional structure without explicit temporal windows containing multiple consecutive monthly observations for each sample. The RF_CV and GB_CV configurations were optimized using RandomizedSearchCV with 10 and 20 parameter combinations, respectively, three cross-validation folds, and a random seed of 42. This experimental design enabled a comparison of the predictive performance of the implemented configurations rather than an assessment of the maximum achievable performance of each algorithm family.

2.8 Temporal split and evaluation metrics

The processed dataset was split chronologically into 455 station-month records from 2017–2024 (88.35%) for model development and 60 records from 2025 (11.65%) for model comparison. Before the split, missing NCC values and other numerical features had been reconstructed using diffusion-inspired imputation followed by iterative imputation based on HistGradientBoostingRegressor. Consequently, both development and evaluation records came from a reference series combining original observations with reconstructed values. Because reconstruction was performed on the broader series before the split, the 2025 comparison should not be interpreted as an independent out-of-time test against new laboratory measurements.

Model performance was assessed using five complementary statistical metrics: the coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), median absolute error (MedAE), and Nash–Sutcliffe efficiency (NSE). R^2 was used to quantify how well the predictions reproduced the variability of NCC in the processed reference dataset, whereas RMSE measured the overall magnitude of prediction errors while assigning greater weight to large deviations. MAE provided the average absolute difference between predicted and reference values, offering a more direct measure of prediction error magnitude. MedAE characterized the typical absolute error and was less sensitive to extreme deviations, which was particularly relevant given the substantial variability in NCC across the monitoring stations. NSE was included as an additional measure of predictive efficiency relative to using the mean reference value as a constant prediction. Under the definitions adopted in this study, R^2 and NSE are mathematically equivalent when calculated using the same reference values and predictions.

All metrics were calculated against NCC in the processed monthly series, combining observed laboratory values and reconstructed estimates. The results therefore describe agreement with that processed reference rather than independent predictive accuracy for laboratory-measured NCC. The same 2025 block was used to compare all 14 configurations, and the model with the highest R^2 was selected for exploratory mapping. This procedure supports comparison among these implementations on this dataset, but does not establish out-of-sample superiority over other algorithms or environmental conditions.

2.9 Spatial application

MLP_Keras was applied to pixels classified as water in the Sentinel-2 scene acquired on December 1, 2025. Spectral predictors were calculated where the required bands were available. Missing predictors, including station-derived temporal features unavailable for single-scene pixels and CIRRUS_INDEX when B10 was absent, were replaced through the median imputer in the prediction pipeline. Thus, the spatial application did not reproduce the full feature information available during station-level model development and must be considered exploratory.

The maps also involve a change in spatial support: station-level predictors were calibrated using means from 3 × 3-pixel windows, whereas spatial estimates were produced for individual pixels. The product depicts relative differences between classes of model-predicted NCC; it is neither a field-validated cell-count map nor a cyanotoxin map.

3 Results

3.1 Data structure and NCC patterns

The regularized series contained 515 station-month records from 2017 to 2025. Before reconstruction, 112 records had available NCC and 403 were missing. The completed matrix allowed numerical comparison across all months but necessarily combined observed and reconstructed reference values. NCC magnitudes also varied markedly among the five stations, particularly at BILL02900, affecting interpretation of aggregate metrics.

3.2 Model comparison

Table 4 summarizes the 14 configurations evaluated against 60 station-month entries in the processed 2025 series. MLP_Keras had the highest R^2 (0.8525), followed by MLP_ScikitLearn (0.7907), PLS (0.7823), Elastic Net (0.7791), and RF_CV (0.7783). MLP_Keras yielded RMSE = 222,998,226.43 cells/mL, MAE = 72,310,835.10 cells/mL, and MedAE = 1,759,164.65 cells/mL. These metrics quantify agreement with a reference that includes reconstructed NCC, rather than accuracy measured exclusively against laboratory observations.

Table 4. Performance of the 14 configurations against the processed monthly reference for 2025 (60 station-month records).

Model	R^2	RMSE (cells mL^{-1})	MAE (cells mL^{-1})	MedAE (cells mL^{-1})
MLP_Keras	0.8525	222,998,226.43	72,310,835.10	1,759,164.65
MLP_ScikitLearn	0.7907	265,620,760.31	88,213,593.82	1,720,036.87

Model	R^2	RMSE (cells mL^{-1})	MAE (cells mL^{-1})	MedAE (cells mL^{-1})
PLS	0.7823	270,909,892.71	98,964,099.27	5,652,354.89
Elastic Net	0.7791	272,885,477.20	98,352,375.28	3,976,383.82
RF_CV	0.7783	273,427,379.20	76,775,350.48	1,417,972.47
Linear regression	0.7648	281,584,028.90	102,655,102.44	4,700,228.10
LSTM	0.7408	295,625,730.01	98,875,788.42	1,281,386.85
RF	0.7398	296,188,900.41	80,655,539.66	1,339,035.58
Extra Trees	0.7271	303,330,015.72	104,780,423.99	1,074,854.54
KNN	0.7095	312,962,922.64	92,153,167.09	701,842.41
GB	0.6335	351,506,952.20	111,641,053.65	937,996.86
GB_CV	0.5573	386,329,246.23	136,024,402.97	1,801,846.82
SVR	−0.1733	628,973,077.84	241,869,218.90	1,499,827.03
GPR	−157.1893	7,303,177,180.31	1,501,397,875.00	3,212,197.73

Note: For the processed reference values, R^2 and NSE have the same algebraic expression and therefore the same numerical value.

The comparison reveals heterogeneous behavior: although MLP_Keras has the highest R^2 and lowest RMSE, RF_CV approaches its MAE, and several models have lower MedAE values. In the evaluated configurations, SVR and GPR yielded negative R^2 values, indicating performance below that of a constant predictor based on the reference mean. These results are specific to the fitted configurations and do not indicate a general limitation of these model families.

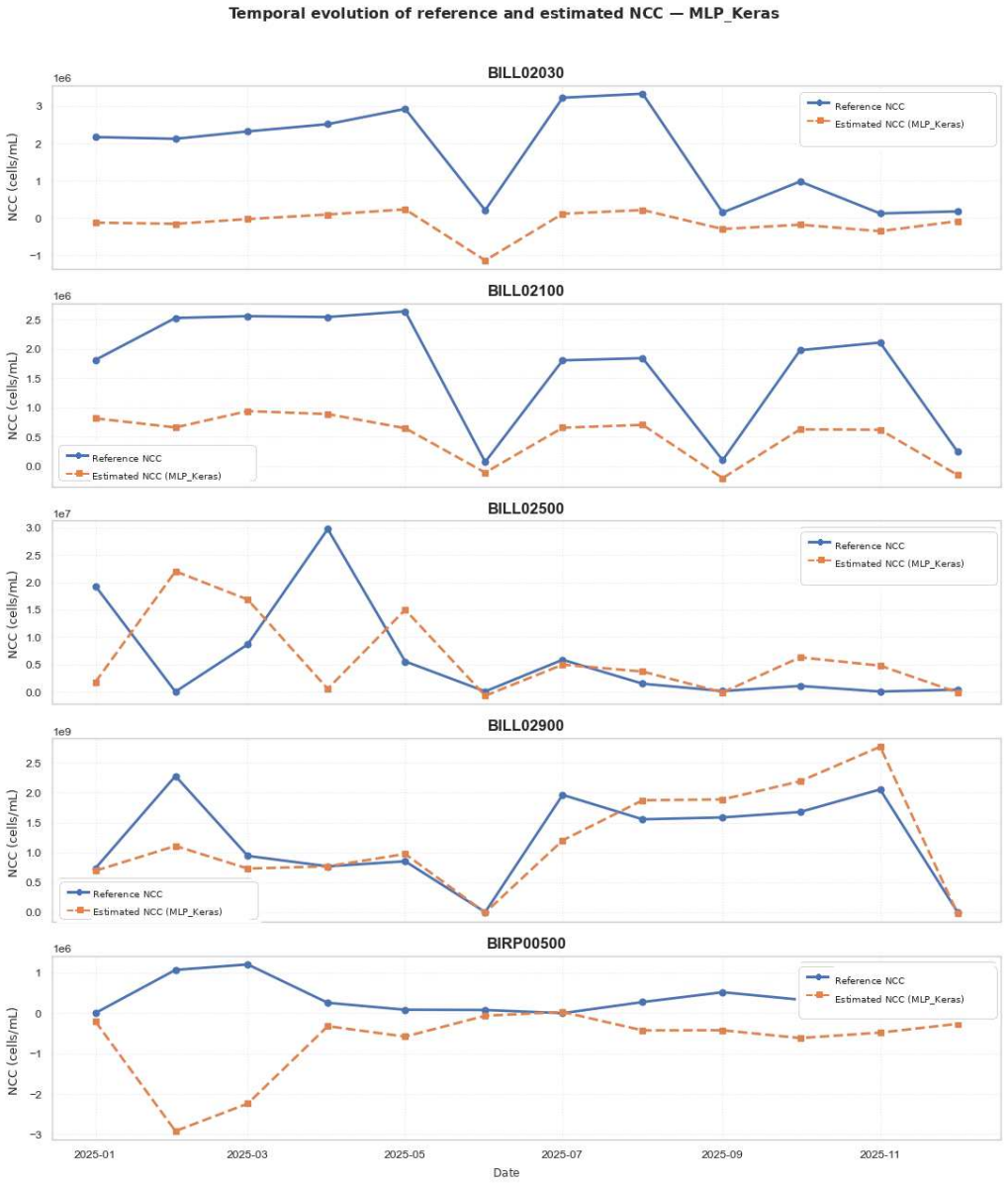


Figure 3. Monthly variation in reference NCC in the processed dataset (including originally observed or reconstructed values) and MLP_Keras predictions during 2025 at the five CETESB monitoring stations.

3.3 Error distribution and spatial interpretation

The MAE of MLP_Keras (72.31 million cells/mL) was approximately 41 times its MedAE (1.76 million cells/mL), while RMSE reached 223 million cells/mL. This substantial difference indicates that large prediction errors strongly influence the aggregate performance metrics. Furthermore, differences in NCC magnitude among monitoring stations, particularly BILL02900, may cause the overall R^2 to be disproportionately influenced by observations with higher values.

Figure 3 illustrates the temporal variation in monthly NCC reference values and the corresponding predictions generated by MLP_Keras for each of the five CETESB monitoring stations throughout 2025. The reference series comprises both originally observed and reconstructed NCC values, whereas the predicted series represents the outputs of the trained model. Visual inspection reveals partial agreement between the temporal patterns of the reference and predicted series at selected monitoring stations. In particular, some decreases and subsequent increases are reflected in the predictions for BILL02100 and BILL02900, although differences in their magnitudes remain evident. At other stations, including BILL02030, BILL02500, and BIRP00500, more pronounced discrepancies are observed, indicating that the correspondence between the predicted and reference temporal patterns varies across monitoring locations.

The predictions also exhibit differences in the magnitude of NCC peaks and, in some months, negative values resulting from the unconstrained linear output layer of the MLP_Keras architecture. Such negative estimates have no physical interpretation as cyanobacterial cell counts. These findings indicate that, although certain temporal variations in the processed NCC reference are reflected in the model outputs, the quantitative agreement is affected by substantial errors and inter-station heterogeneity.

Overall, Figure 3 provides a complementary visual assessment of the model's behavior across the monitoring stations, highlighting both similarities in selected temporal patterns and limitations in reproducing NCC magnitudes. Because the reference series includes reconstructed values and the reconstruction procedure preceded the chronological data split, this comparison should be interpreted as an exploratory assessment of agreement with the processed monthly series rather than independent validation against laboratory measurements.

3.4 Spatial estimation

Figure 4 displays the MLP_Keras output for pixels classified as water in the Sentinel-2 scene of December 1, 2025. Estimated NCC values were grouped into the descriptive classes A–D to visualize spatial contrasts associated with the predictors available at pixel level. Because several predictors were replaced by imputed values during spatial inference, these maps are exploratory numerical model outputs rather than field-validated estimates of cell abundance across the reservoir.

Environmental interpretation of the spatial estimates of the Number of Cyanobacterial Cells (NCC) considered monitoring thresholds established by the Brazilian Ministry of Health and technical guidance available in the literature. Ordinance GM/MS No. 888/2021 requires weekly cyanobacteria monitoring at surface-water intake points used for public water supply when counts exceed 10,000 cells/mL, and cyanotoxin analyses at least weekly when counts exceed 20,000 cells/mL (BRASIL, 2021).

Historically, the Ministry of Health also proposed surveillance and alert levels for tracking blooms, using indicative ranges of 10,000–20,000 cells/mL, 20,000–100,000 cells/mL, and values above 100,000 cells/mL, associated with different stages of bloom development and monitoring needs (BRASIL, 2003). These references are complemented by monitoring guidance from the Ministry of Health (BRASIL, 2016) and the CETESB technical manual (2013).

Based on these references, NCC estimates were divided into four cartographic classes, described in Table 5 and shown in Figure 4.

Table 5. Cartographic ranges of estimated NCC and their environmental interpretation.

NCC RANGE (CELLS/ML)	ENVIRONMENTAL INTERPRETATION
A — BELOW 20,000	Density below the threshold for intensified cyanotoxin analyses. Includes values above 10,000 cells/mL, associated with more frequent cyanobacteria monitoring.
B — 20,000 TO <50,000	Density warranting attention, close to or above the 20,000 cells/mL reference. Values strictly above this threshold, when measured at intake points, require weekly cyanotoxin analyses.
C — 50,000 TO <100,000	High cyanobacterial density, potentially associated with increased proliferation. The 50,000 cells/mL boundary is a cartographic subdivision, not a specific regulatory threshold.
D — 100,000 OR HIGHER	Very high density, consistent with conditions requiring special attention to possible blooms. Historical technical references support this value, but it does not by itself confirm a toxic bloom.

The descriptive classes permit examination of spatial differences in model-estimated NCC across the reservoir. Boundaries at 50,000 and 100,000 cells/mL distinguish higher numerical ranges, while the reference values of 10,000 and 20,000 cells/mL provide context from technical monitoring guidance. The classes do not establish compliance with regulatory monitoring criteria.

These classes are intended solely for descriptive and exploratory purposes. Regulatory criteria apply to counts actually measured at water intake points, not to estimates derived from remote sensing. Furthermore, cell counts alone do not determine the presence of cyanotoxins, the toxicity of the species present, or actual public health risk, which require additional information and specific analyses.

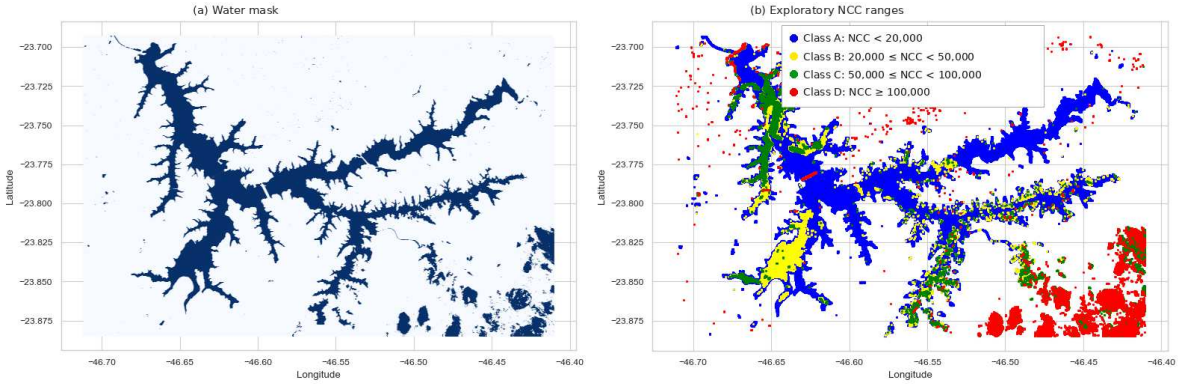


Figure 4. Billings Reservoir: (a) water-surface mask and (b) exploratory spatial distribution of estimated NCC for the Sentinel-2 scene of December 1, 2025, grouped into four cell-density ranges. The intervals are descriptive classes of estimated concentration, not official public health risk categories.

4 Discussion

4.1 Spectral interpretation and predictor relevance

Eight features were included in MLP_Keras: FAI_pct, ulyssys, SABI_z, CIrededge_pct, MCI, CIRRUS_INDEX, TURB_R_G_z, and NDCI_delta. NDCI and CIrededge exploit red/red-edge contrasts relevant to productive waters (Mishra and Mishra 2012; Gitelson et al. 2009); MCI quantifies excess spectral response near 705 nm and is used to characterize algal blooms (Binding et al. 2013); and FAI highlights NIR reflectance above a spectral baseline (Hu 2009). SABI combines visible and NIR bands to distinguish surface algal material (Alawadi 2010). TURB_R_G, meanwhile, is an optical ratio related to constituents that modify the red/green response.

The indices summarize relationships between bands and introduce interpretable information into the imputation and regression processes. Under specific conditions, spectral differences and ratios help reduce common sensitivities to illumination or radiometric scaling, although they do not eliminate imperfect atmospheric correction, adjacency effects, or measurement noise. In the present workflow, indices, bands, and temporal variables were reconstructed jointly; accordingly, using indices before imputation should be understood as a physically grounded feature engineering choice, not as an experimental comparison demonstrating superiority over direct band imputation.

The UWQV product is derived from Sentinel-2 observations and relates to optical properties of water, including chlorophyll-associated responses and suspended material (Zlinszky and Padányi-Gulyás 2020). Here, ulyssys represents the first band of an exported TIFF used as a numerical input; without confirmed information about the specific export, no independent physical unit can be assigned to it. Likewise, because B10 is generally unavailable in Level-2A products, imputed CIRRUS_INDEX values cannot be interpreted as direct cirrus-band observations.

4.2 Predictive performance and interpretation of the temporal test

On the processed 2025 block, MLP_Keras reproduced variation in the reconstructed monthly NCC reference. $R^2 = 0.8525$ quantifies this within-pipeline agreement and must not be interpreted as validation against 60 independent laboratory samples. This distinction is essential when comparing the present metrics with those of studies evaluated exclusively on measured cyanobacterial cell counts.

The high proportion of missing values in the monthly series makes reconstruction a central component of the experiment. Correlations between features and the processed target may reflect both environmental relationships and patterns introduced by imputation, particularly during extreme events. In this context, RMSE, MAE, and MedAE offer complementary perspectives: RMSE is more sensitive to large errors, whereas MedAE describes the central tendency of the error distribution.

The chronological 2017–2024/2025 partition was introduced after the monthly reconstruction step. Consequently, the regressors were compared over a later calendar period, but their target reference had been generated using information from the broader time series. This is a retrospective comparison on processed data and does not rule out information leakage from the reconstruction stage.

4.3 Comparison of model architectures

MLP_Keras achieved a higher overall R^2 than MLP_ScikitLearn, PLS, and the other compared models. Differences in architecture, optimization, preprocessing, and early stopping may explain some of the variation, but the improvement cannot be attributed to any single component. The inputs used in the supervised LSTM configuration did not represent long

sequential windows; its results therefore characterize the performance of that specific implementation rather than conclusively assessing the benefits of temporal recurrence.

The SVR and GPR configurations produced R^2 values below zero, with GPR especially sensitive to deviations on the NCC scale. This finding reinforces that algorithm choice depends on the target distribution, hyperparameter settings, and feature processing, without implying that these methods are generally unsuitable for remote sensing.

4.4 Applicability and limitations

The workflow integrated CETESB observations from five stations with Sentinel-2 imagery, generated monthly series and presented an exploratory spatial visualization. Its interpretation is constrained by the number of stations, the large fraction of reconstructed NCC values, the pre-split reconstruction procedure, and the change from 3 × 3-pixel extraction windows to individual-pixel mapping. The outputs should therefore be regarded as methodological and exploratory rather than an operational monitoring product.

Sources of uncertainty include the possible seven-day mismatch between field sampling and image acquisition, the reservoir's optical heterogeneity, the use of indices sensitive to multiple water constituents, reconstruction of bands, indices, and NCC, the unavailability of certain predictors for individual pixels, and the influence of very large cell counts. Map interpretation must also recognize that estimated values do not replace laboratory-based cyanobacteria and cyanotoxin monitoring protocols.

The 20,000 cells/mL reference was used to contextualize the exploratory cartographic classes. Under Ordinance GM/MS No. 888/2021, measured counts strictly exceeding this value at relevant water-intake points trigger at least weekly cyanotoxin analyses; the distinct threshold of 10,000 cells/mL pertains to cyanobacterial monitoring frequency. Neither requirement can be inferred from satellite-derived class assignments, which do not establish toxicity or regulatory status.

4.5 Comparison with the state of the art

Pompêo et al. (2021) provide a direct precedent for using Sentinel-2 to investigate NCC in reservoirs of São Paulo State. The contribution of the present research is a Billings-specific implementation that integrates spectral and temporal descriptors, monthly reconstruction, and comparison of regression configurations, complemented by a single-scene spatial demonstration. Table 6 places this implementation alongside selected studies while acknowledging differences in their target variables and evaluation protocols.

Davis et al. (2023) compared machine learning algorithms for observing blooms in Utah Lake; Dorne et al. (2024) introduced CyFi for estimates across multiple water bodies; and Ge et al. (2024) investigated regression alternatives using Sentinel-2 data and fishpond monitoring. Differences in scale, optical conditions, experimental reference, and validation objectives prevent direct ranking of these studies based on R^2 alone.

Taken together, the literature indicates that spectral indices and algorithms can support observation of bloom-related phenomena, but numerical performance depends on environmental characteristics and the available reference data. This study adds an integrated description of data processing and modeling for a Brazilian urban reservoir, including a spatial demonstration and quantitative evaluation against the processed monthly series.

The diversity of applications underscores the importance of interpreting estimates in light of hydrological context, spatial scale, and the quality of records used for calibration.

Table 6. Contextual comparison of the approach developed here with representative studies from the literature.

Study / approach	Data and target variable	Validation / reported outcome	Relationship to this study
Davis et al. (2023), Utah Lake, USA. Sentinel-2, feature selection, and comparison of four regression algorithms.	In situ samples and Sentinel-2 imagery; cyanobacterial concentration.	Four-model comparison using different error criteria and a protocol and scale distinct from those used here.	Comparable spectral selection approach; this study used the processed monthly series as its reference.
Dorne et al. (2024), CyFi. Gradient-boosted trees (LightGBM), open regional tool.	Sentinel-2 and in situ cyanobacterial density measurements in US inland waters.	Multi-regional evaluation; density estimation and severity classification.	CyFi emphasizes geographic generalization; this study calibrates continuous regression locally for the Billings Reservoir.
Ge et al. (2024), Třeboň fishponds, Czech Republic. PLSR, Random Forest, SVM, and XGBoost.	740 samples; Sentinel-2 and AlgaeTorch; cyanobacteria and related variables.	SVM: $R^2 = 0.88$ in single-pond validation and $R^2 = 0.56$ in joint validation.	Cross-domain performance loss reinforces the importance of spatial validation before model transfer.
Pompêo et al. (2021), Cantareira System reservoirs, Brazil. Water quality parameter estimation using Sentinel-2.	NCC, chlorophyll-a, and Secchi disk depth in São Paulo reservoirs.	Empirical modeling based on reference measurements; design and metrics are not directly interchangeable with the 2025 test.	Direct precedent for Sentinel-2-based NCC estimation; the present study examines Billings using multivariate imputation and internal model comparison.

Study / approach	Data and target variable	Validation / reported outcome	Relationship to this study
This study, Billings, Brazil: 14 configurations, including MLPs, tree models, regression methods, and LSTM.	515 station-month records, with 403 NCC gaps before imputation; five CETESB stations and Sentinel-2 L2A scenes (2017–2025).	2025 evaluation (n = 60 station-months from the reconstructed series): MLP_Keras R^2 = 0.8525; RMSE = 222,998,226.43 and MAE = 72,310,835.10 cells/mL. Metrics use the processed monthly reference, including reconstructed NCC values.	Integrates 3 × 3 extraction, indices, multivariate reconstruction, model comparison, and exploratory mapping, with a monthly reference combining observed and reconstructed values.

Note: Metrics, target variables, reference datasets, and protocols differ among studies; the reported values do not constitute a directly comparable ranking.

4.6 Methodological perspectives

Additional work should assess imputation accuracy independently of the supervised prediction task, compare reconstruction strategies, and evaluate the models exclusively against retained laboratory observations. Other priorities include spatial transfer between stations, performance during high-NCC events, sensitivity to sampling–image timing, and the robustness of pixel-scale predictors. These tests would be needed before the workflow could support operational interpretation.

5 Conclusions

This study integrated CETESB monitoring data and Sentinel-2 scenes for the Billings Reservoir, using extraction from 3 × 3-pixel windows, temporal matching within seven days, and indices related to water, pigments, surface biomass, and turbidity. Monthly expansion yielded 515 station-month records, of which 403 lacked NCC before multivariate imputation. The reconstruction workflow incorporated bands, indices, and temporal variables and provided the matrix used to compare 14 supervised configurations.

In the processed 2025 reference block, MLP_Keras achieved the highest R^2 (0.8525), with RMSE of 222,998,226.43 and MAE of 72,310,835.10 cells/mL. The difference between MAE and MedAE reveals strong sensitivity to large errors and inter-station heterogeneity. Because the reference incorporates reconstructed NCC and reconstruction preceded temporal splitting, these metrics express agreement with the processed series and do not establish predictive performance against independent laboratory measurements.

The December 2025 map illustrates a single-scene application to water-classified pixels and depicts relative ranges of model-estimated NCC. The mapped classes are descriptive: the 20,000 cells/mL regulatory reference concerns measured counts at relevant intake points and is not a satellite-based determination of cyanotoxin presence. Overall, the work contributes a documented integration of monitoring records, spectral–temporal features, reconstruction, and model comparison, while identifying the validation steps needed for any future operational application.

References

Alawadi, F.: Detection of surface algal blooms using the newly developed algorithm surface algal bloom index (SABI). In: Remote Sensing of the Ocean, Sea Ice, and Large Water Regions 2010, Proc. SPIE 7825 (2010). https://doi.org/10.1117/12.862096

Binding, C.E., Greenberg, T.A., Bukata, R.P.: The MERIS maximum chlorophyll index; its merits and limitations for inland water algal bloom monitoring. Journal of Great Lakes Research 39, 100–107 (2013).

BRASIL. Ministry of Health. National Health Foundation. Cianobactérias tóxicas na água para consumo humano: impactos na saúde pública e processos de remoção em água para consumo humano [Toxic cyanobacteria in drinking water: public health impacts and removal processes]. Brasília: National Health Foundation, 2003. Available at: https://bvsms.saude.gov.br/bvs/publicacoes/mnl_ciano_bacterias.pdf. Accessed: 17 Oct. 2026.

BRASIL. Ministry of Health. Health Surveillance Secretariat: Orientações técnicas para o monitoramento de cianobactérias/cianotoxinas nos mananciais de abastecimento de água para consumo humano [Technical guidance for monitoring cyanobacteria/cyanotoxins in drinking-water sources]. Ministry of Health, Brasília (2016). https://bvsms.saude.gov.br/bvs/publicacoes/orientacoes_tecnicas_cianobacterias_cianotoxinas_agua.pdf (accessed 13 Oct. 2026).

BRASIL. Ministry of Health: Ordinance GM/MS No. 888 of May 4, 2021. Amends Annex XX of Consolidation Ordinance No. 5/GM/MS of September 28, 2017. Brasília, DF (2021). https://bvsms.saude.gov.br/bvs/saudelegis/gm/2021/prt0888_24_05_2021_rep.html (accessed 13 Oct. 2026).

Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

CETESB—São Paulo State Environmental Agency: Boletim de qualidade das águas brutas do Sistema Billings [Billings System raw water quality bulletin] (n.d.). https://cetesb.sp.gov.br/aguas-interiores/ (accessed 16 May 2026).

CETESB—São Paulo State Environmental Agency: Manual de cianobactérias planctônicas: legislação, orientações para o monitoramento e aspectos ambientais [Manual of planktonic cyanobacteria: legislation, monitoring guidance, and environmental aspects]. CETESB, São Paulo (2013). https://repositorio.cetesb.sp.gov.br/handle/123456789/2208 (accessed 13 Oct. 2026).

Codd, G.A.: Cyanobacterial toxins, the perception of water quality, and the prioritisation of eutrophication control. Ecological Engineering 16, 51–60 (2000). https://doi.org/10.1016/S0925-8574(00)00089-6

Davis, R., Panja, P., McLennan, J.: Integrated workflow for interpretation of satellite imageries using machine learning to assess and monitor algal blooms in Utah Lake, USA. Ecological Informatics 75, 102033 (2023). https://doi.org/10.1016/j.ecoinf.2023.102033

Dorne, E., Wetstone, K., Cerquera, T.B., Gupta, S.: Cyanobacteria detection in small, inland water bodies with CyFi. In: Proceedings of the 23rd Python in Science Conference, pp. 154–173 (2024). https://doi.org/10.25080/PDHK7238

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., et al.: Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sensing of Environment 120, 25–36 (2012). https://doi.org/10.1016/j.rse.2011.11.026

European Space Agency: Sentinel-2 Level-2A Algorithm Theoretical Basis Document. ESA (2021).

Feyisa, G.L., Meilby, H., Fensholt, R., Proud, S.R.: Automated water extraction index: a new technique for surface water mapping using Landsat imagery. Remote Sensing of Environment 140, 23–35 (2014). https://doi.org/10.1016/j.rse.2013.08.029

Fickas, K.C., O'Shea, R.E., Pahlevan, N., Smith, B., Bartlett, S.L., Wolny, J.L.: Leveraging multimission satellite data for spatiotemporally coherent cyanoHAB monitoring. Frontiers in Remote Sensing 4, 1157609 (2023). https://doi.org/10.3389/frsen.2023.1157609

Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research 20, 1–81 (2019).

Friedman, J.H.: Greedy function approximation: a gradient boosting machine. The Annals of Statistics 29, 1189–1232 (2001).

Gao, B.C.: NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment 58, 257–266 (1996). https://doi.org/10.1016/S0034-4257(96)00067-3

Ge, Y., Shen, F., Sklenička, P., Vymazal, J., Baxa, M., Chen, Z.: Machine learning for cyanobacteria inversion via remote sensing and AlgaeTorch in the Třeboň fishponds, Czech Republic. Science of the Total Environment 947, 174504 (2024). https://doi.org/10.1016/j.scitotenv.2024.174504

Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning 63, 3–42 (2006).

Giardino, C., Brando, V.E., Dekker, A.G., Strömbeck, N., Candiani, G.: Assessment of water quality in Lake Garda using Hyperion. Remote Sensing of Environment 109, 183–195 (2007). https://doi.org/10.1016/j.rse.2006.12.017

Gitelson, A.A., Gurlin, D., Moses, W., Barrow, T.: A bio-optical algorithm for the remote estimation of chlorophyll-a concentration in case 2 waters. Environmental Research Letters 4, 045003 (2009). https://doi.org/10.1088/1748-9326/4/4/045003

Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2009).

Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9, 1735–1780 (1997).

Hu, C.: A novel ocean color index to detect floating algae in the global oceans. Remote Sensing of Environment 113, 2118–2129 (2009). https://doi.org/10.1016/j.rse.2009.05.012

Kuhn, C., Valerio, G., Bernard, C., Hubert, P., Coquery, M., Faucheux, M., Lavaux, T.: Water constituents and water depth retrieval from Sentinel-2A—A first evaluation in an oligotrophic lake. Remote Sensing 8, 941 (2016). https://doi.org/10.3390/rs8110941

Lamparelli, M.C., Carvalho, M.C., Espindola, E.L.G.: Reservoir eutrophication in São Paulo State: evaluation and monitoring perspectives. Brazilian Journal of Biology 68, 607–618 (2008). https://doi.org/10.1590/S1519-69842008000300020

Llodrà-Llabrés, J., Martínez-López, J., Postma, T., Pérez-Martínez, C., Alcaraz-Segura, D.: Retrieving water chlorophyll-a concentration in inland waters from Sentinel-2 imagery: Review of operability, performance and ways forward. International Journal of Applied Earth Observation and Geoinformation 125, 103605 (2023). https://doi.org/10.1016/j.jag.2023.103605

Matthews, M.W.: A current review of empirical procedures of remote sensing in inland and near-coastal transitional waters. International Journal of Remote Sensing 32, 6855–6899 (2011).

McFeeters, S.K.: The use of the normalized difference water index (NDWI) in the delineation of open water features. International Journal of Remote Sensing 17, 1425–1432 (1996). https://doi.org/10.1080/01431169608948714

Mishra, S., Mishra, D.R.: Normalized difference chlorophyll index: a novel model for remote estimation of chlorophyll-a concentration in turbid productive waters. Remote Sensing of Environment 117, 394–406 (2012). https://doi.org/10.1016/j.rse.2011.10.016

O'Brien, R.M.: A caution regarding rules of thumb for variance inflation factors. Quality & Quantity 41, 673–690 (2007). https://doi.org/10.1007/s11135-006-9018-6

O'Neil, J.M., Davis, T.W., Burford, M.A., Gobler, C.J.: The rise of harmful cyanobacteria blooms: the potential roles of eutrophication and climate change. Harmful Algae 14, 313–334 (2012). https://doi.org/10.1016/j.hal.2011.10.027

Pearson, K.: Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London 58, 240–242 (1895). https://doi.org/10.1098/rspl.1895.0041

Pompêo, M., Moschini-Carlos, V., Bitencourt, M.D., Sòria-Perpinyà, X., Vicente, E., Delegido, J.: Water quality assessment using Sentinel-2 imagery with estimates of chlorophyll a, Secchi disk depth, and cyanobacteria cell number: the Cantareira System reservoirs (São Paulo, Brazil). Environmental Science and Pollution Research 28, 34990–35011 (2021). https://doi.org/10.1007/s11356-021-12975-x

Pérez-González, R., Urrego, P., Delegido, J., Soria-Perpinyà, X., Ruiz-Verdú, A., Vicente, E., Moreno, J.: Phycocyanin monitoring in inland water bodies using Sentinel-2 imagery. Water 13, 2866 (2021). https://doi.org/10.3390/w13202866

Reichstein, M., Camps-Valls, G., Stevens, B., et al.: Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204 (2019).

Sentinel Hub: Ulyssys Water Quality Viewer (UWQV). Custom Scripts for Sentinel-2 (n.d.). https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/ulyssys_water_quality_viewer/ (accessed 16 Oct. 2026).

Shi, K., Zhang, Y., Qin, B., Zhou, B.: Remote sensing of cyanobacterial blooms in inland waters: present knowledge and future challenges. Science Bulletin 64(20), 1540–1556 (2019). https://doi.org/10.1016/j.scib.2019.07.002

Tashiro, Y., Song, J., Song, Y., Ermon, S.: CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In: Advances in Neural Information Processing Systems 34 (2021). https://proceedings.neurips.cc/paper/2021/hash/cfe8504bda37b575c70ee1a8276f3486-Abstract.html

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 13, 8570 (2021). https://doi.org/10.3390/su13158570

Xu, H.: Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. International Journal of Remote Sensing 27, 3025–3033 (2006). https://doi.org/10.1080/01431160600589179

Zlinszky, A., Padányi-Gulyás, G.: Ulyssys Water Quality Viewer Technical Description Supplementary. Preprints 2020, 2020010386 (2020). https://doi.org/10.20944/preprints202001.0386.v1

Statements and Declarations

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Data availability — Monitoring data were obtained from CETESB records, and satellite information came from Sentinel-2 Level-2A products. The resulting analytical dataset consists of processed monthly records with spectral indicators and observed or reconstructed NCC values, as described in Materials and methods.

Code availability — Processing was developed in Python notebooks covering multispectral analysis, multivariate time-series reconstruction, model comparison, and spatial mapping of predictions.