

Article

Not peer-reviewed version

Evaluating Three Machine Learning Methods for Simulating Chlorophyll-a in Data-Scarce Lakes of Leon County, Florida

Alexander Wang and [Ming Ye](#)*

Posted Date: 14 September 2026

doi: 10.20944/preprints202609.1068.v1

Keywords: Random Forest; XGBoost; LSTM; water quality; temporal and spatial validation; method-agnostic importance analysis



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

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.

Article

Evaluating Three Machine Learning Methods for Simulating Chlorophyll-a in Data-Scarce Lakes of Leon County, Florida

Alexander Wang ¹ and Ming Ye ^{2,*}

¹ College of Science, Georgia Institute of Technology, Atlanta, GA 30332, USA

² Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, FL 32306, USA

* Correspondence: mye@fsu.edu

Abstract

Chlorophyll-a (Chl-a), the primary photosynthetic pigment in algae, is widely used as a proxy for phytoplankton biomass and trophic status in lakes. While machine learning (ML) methods have been used for simulating Chl-a concentrations, there have been few studies examining ML methods for lakes with scarce monitoring data of water quality. This study evaluated simulation performance of three commonly used ML methods (i.e., Random Forest, XGBoost, and a Long Short-Term Memory (LSTM) network) for 11 data-scarce lakes in Leon County, FL. Each lake has only 1 – 6 sampling locations with a quarterly monitoring schedule, and there are a total of 1,012 sets of observations (43 to 252 per lake) for the study period of 2014 – 2024. Each observation set includes 43 water quality variables, e.g., phosphorus, organic nitrogen, and dissolved oxygen. Evaluating the three ML methods was based on three cross-validation operations and three statistical metrics (i.e., root mean square error, mean absolute error, and coefficient of determination). In a five-fold cross-validation, XGBoost achieved the best simulation performance. In a chronological cross-validation, all the methods had a similar modest simulation performance. A spatial cross-validation revealed that it was challenging to simulate Chl-a concentrations for two lakes with high Chl-a concentrations and large temporal variations. A method-agnostic importance test consistently identified a subset of water quality variables as top predictors across all the methods, including phosphorus, organic nitrogen, and biochemical oxygen demand. These findings demonstrate that ML is an effective tool for simulating Chl-a concentrations even for data-scarce lakes, whereas more monitoring data are needed for simulating high Chl-a concentrations with large temporal variations.

Keywords: Random Forest; XGBoost; LSTM; water quality; temporal and spatial validation; method-agnostic importance analysis

1. Introduction

Eutrophication is a persistent water quality concern because excess nutrient enrichment can increase phytoplankton biomass, reduce dissolved oxygen, alter aquatic food webs, impair recreation, and increase the likelihood of harmful algal blooms [1–4]. Chlorophyll-a (Chl-a), the dominant photosynthetic pigment measured in algae and cyanobacteria, is widely used as a practical indicator of algal biomass and trophic state [5,6]. Machine learning (ML) methods have become popular for simulating Chl-a concentrations, because they can represent complex nonlinear relations among water quality variables (e.g., Chl-a, nutrients, optical conditions, oxygen, temperature, and pH) [9–12]. Data availability for ML-based water quality modeling can vary by more than an order of magnitude depending on the monitoring approaches and study objectives. Table 1 summarizes the number of data and data sources used in several recent literature on ML-based water quality modeling. Chen et al. [19] compared 10 machine-learning methods using 33,612 observations

collected from major rivers and lakes in China during 2012–2018. Zhao et al. [36] analyzed the nutrient–Chl-a relationship using 25,083 observations from 2,849 lakes worldwide. Pahlevan et al. [37] developed a machine-learning approach for Chl-a retrieval using 2,943 matched field and remote-sensing observations. At an even larger scale, Zhi et al. [31] used the CAMELS-Chem database for 236 minimally disturbed watersheds; a recent compilation of U.S. lake and stream chlorophyll-a measurements contains more than 1.37 million pigment measurements from approximately 84,000 sites collected during 2005–2022 and was developed specifically to support machine-learning, remote-sensing, and process-based modeling applications [38]. Other studies have also used remote-sensing observations that provide repeated spatial coverage or high-frequency measurements that are not available from conventional grab-sampling programs [7,13,14]. These studies demonstrate that a substantial portion of water-quality ML research benefits from large regional, national, or globally distributed datasets with thousands to millions of observations.

Table 1. Summary of data used in several literature for evaluating machine learning methods for water quality and Chl-a simulations.

Reference	Number of Data	Data Source	Study Application
Chen et al. (2020) [19]	33,612	Major rivers and lakes in China, 2012–2018	Surface water water quality prediction
Zhao et al. (2023) [36]	25,083	Global dataset covering 2,849 lakes worldwide	Nutrient-Chl-a relationships across lakes
Pahlevan et al. (2020) [37]	2,943	Satellite/field matchup observations across diverse water bodies	Chl-a retrieval in inland and coastal waters
Zhi et al. (2021) [31]	Tens of thousands of data	Large-scale river monitoring observations across United States	Continental-scale water-quality prediction
Spaulding et al. (2024) [38]	1,374,000	United States, 2005–2022	Chl-a monitoring across U.S. lakes and streams
Figueiredo et al. (2026) [39]	82	Matched hyperspectral-Chl-a observations from Pampulha Reservoir, Brazil	Chl-a retrieval in an optically complex urban reservoir
Cubilla-Montilla et al. (2026) [40]	1,176	Monthly observations from 14 stations in Lake Gatun, Panama, 2017-2023	Chl-a prediction and predictor interpretation

However, for many lakes, their monitoring programs commonly collect monthly or quarterly grab samples at a limited number of sampling locations, resulting in relatively small datasets with irregular sampling intervals, missing predictor variables, and uneven representation among lakes. Recent Limnological Review studies illustrate the range of sample sizes in site-specific applications: Figueiredo et al. [39] evaluated hyperspectral Chl-a retrieval using 82 sets of water quality monitoring data from a single reservoir, whereas Cubilla-Montilla et al. [40] analyzed 1,176 monthly observations

collected at 14 stations in Lake Gatun. These examples show that ML can be informative for individual water bodies. With the differences in amount of data used for ML modeling, there is a **practical modeling question**: can data-driven ML methods accurately estimate Chl-a concentrations for data-scarce lakes? This question has not been well studied or reported in literature of evaluating ML methods.

This study addressed the above ML modeling question for 11 data-scarce lakes in Leon County, FL. (Figure 1), where each lake has only 1 – 6 sampling locations with quarterly monitoring schedules. The Leon County dataset represents a constrained but common monitoring setting for lakes. For the dataset used in this study, it contains 1,012 station-date observations across 11 lakes, and each lake contributes only 43–252 observations over the 2014–2024 study period. Thus, an individual lake may have only approximately four sampling periods per year at a small number of locations, rather than continuous or daily observations. This distinction is important because the 1,012 observations are distributed across multiple lakes, locations, and years and therefore do not represent 1,012 independent observations of a single lake. The resulting lake-specific sample size is small, and provides a limited representation of short-term Chl-a events with high concentrations.

Unlike published studies that emphasize algorithm development and evaluation using large monitoring datasets, this study evaluated whether commonly used ML methods remain effective for data-scarce lakes. The ML methods considered in this study are two ensemble tree methods, i.e., Random Forest and XGBoost, which are common choices for tabular environmental datasets because they capture interactions of water quality variables and provide interpretable feature rankings [19–24]. This study also considered a Long Short-Term Memory (LSTM) network, which has been widely used in hydrology and environmental studies where temporal dependence of Chl-a on water quality is strong [25–29]. By emphasizing realistic monitoring conditions rather than idealized data-rich environments, this study evaluated the strengths and limitations of the three ML methods for data-scarce lakes.

The evaluation of the three ML methods was based on cross-validation and paid special attention to common problems of cross-validation approaches to avoid overestimating simulation performance. For example, random training-testing splits may place similar observations from the same stations or sampling periods into both the training and testing datasets, resulting in overestimates of simulation performance of ML methods. Similarly, fitting time-series plots may demonstrate interpolation rather than forecasting ability. To address these problems, five-fold cross-validation, chronological cross-validation, and spatial cross-validation were used in this study. The three ML methods were evaluated with the same preprocessing, five-fold cross-validation split, chronological train-test split, and leave-one-lake-out test.

This study addressed the following three specific questions related to evaluating ML methods: (1) Which of the three ML methods has the best simulation performance? (2) How much of the performance is retained when a method trained on data of 2014 – 2019 is tested on data of 2020 – 2024? (3) How well do the methods transfer to a lake withheld entirely from training? The first question was addressed using a five-fold cross-validation, the second question using a chronological cross-validation, and the third question using a spatial (or leave-one-lake-out) cross-validation. This study also identified the water-quality variables most consistently important to Chl-a concentration across the three methods and discussed the practical implications of these variables for lake monitoring.

2. Materials and Methods

2.1. Lake Water Quality Data and Preprocessing

Leon County, FL. (Figure 1) monitors water quality of 11 lakes on a quarterly basis, although sampling frequency varies among lakes and years according to management priorities. The monitoring data for the period 2014–2024 were used in this study. The raw monitoring datasets were processed to exclude all Chl-a and pheophytin variables from the predictor set used by the three ML

methods. This is important because pigment data can be chemically and/or ecologically related to Chl-a and may inflate simulation performance of the ML methods. Following the screening procedure shown in Figure 2, the predictor set includes a total of 43 water quality variables, e.g., nutrients, organic carbon, oxygen, pH, temperature, turbidity, suspended solids, conductance, alkalinity, hardness-related ions, metals, and microbial indicators. After date parsing, numeric conversion, and removal of records without valid dates, station names, characteristic names, or numeric results, the measurements were pivoted to a station-date matrix. Duplicate measurements for the same station, date, and characteristic were averaged. Missing predictor values were filled with training-set medians. The data processing yielded a dataset of 1,012 station-date records, each including the 43 water quality variables used as predictors.

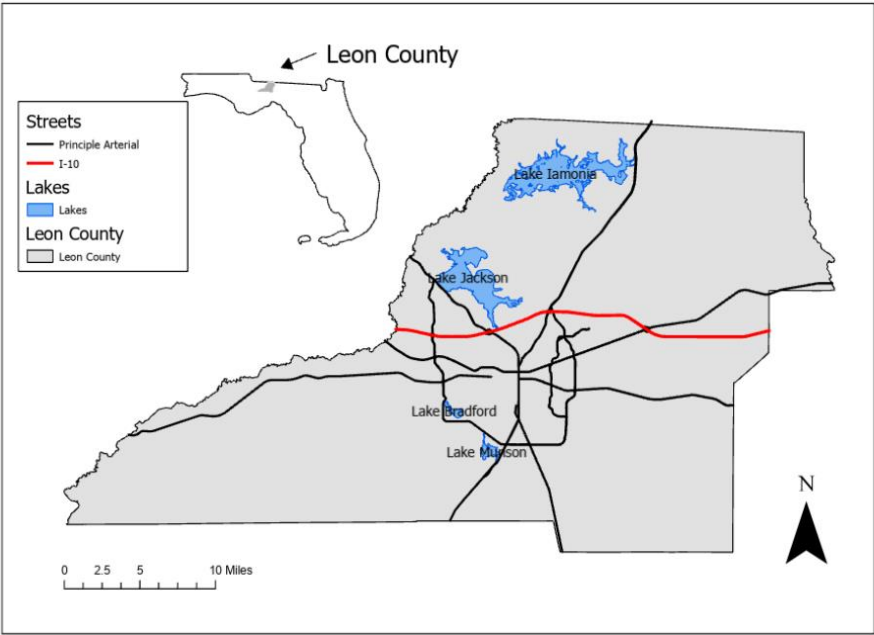


Figure 1. Location of Leon County, FL, and the four lakes (Lake Iamonia, Lake Jackson, Lake Munson, and Lake Bradford) selected for spatial evaluation of ML methods.

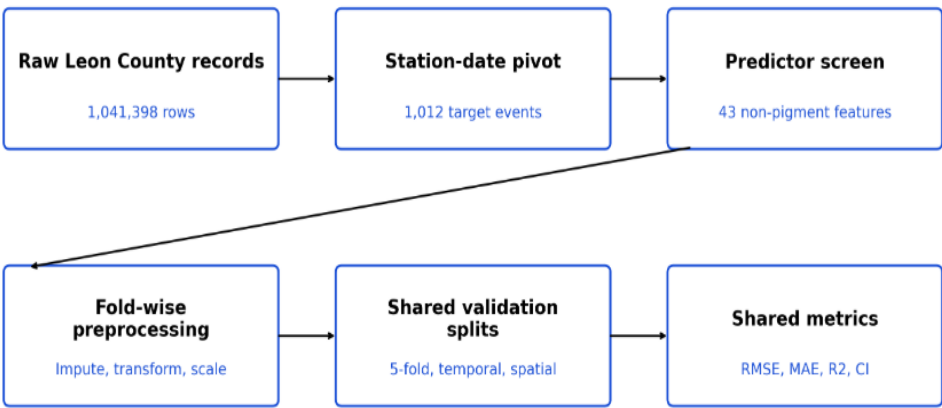


Figure 2. Flowchart of preprocessing water quality data and evaluation of ML methods.

Among the 11 lakes in Leon County, four lakes were selected for the spatial cross-validation, and they are Lake Iamonia, Lake Jackson, Lake Munson, and Lake Bradford with their locations shown in Figure 1. Table 2 summarizes the numbers of sampling locations, number of monitoring data, sampling periods, and basic statistics of Chl-a concentrations for the four lakes. The table indicates data scarcity. For example, each lake has 43 – 252 datasets for the monitoring period of 2014

– 2024. The four selected lakes were also used as an example to demonstrate that the Leon County lakes have a broad range of watershed characteristics, trophic conditions, and hydrologic settings. Lake Iamonia and Lake Jackson, located in northern Leon County, are larger lakes influenced by fluctuating hydrology and wetland connectivity [33,34]. Lake Bradford and Lake Munson, located downstream of the City of Tallahassee, are smaller lakes surrounded by mixed land uses. Lake Iamonia is a rural lake that receives surface runoff from surrounding forests. In contrast, Lake Munson is an urban lake, receiving approximately one third of urban runoff from the City of Tallahassee. The lake has a long history of nutrient enrichment and stormwater influence [32]. Lake Jackson is an aquatic preserve, and Lake Bradford has been used by Florida State University for student recreation.

Table 2. A summary of number of sampling locations, number of datasets, sampling periods, and basic statistics of Chl-a concentrations for Lakes Iamonia, Jackson, Munson, and Bradford.

Lake	Sampling Locations	Data Number	First date	Last date	Mean	Median	Max	Min
Lake Iamonia	6	252	2014-2-20	2024-10-30	5.66	2.90	221.00	1.0
Lake Jackson	4	123	2014-2-19	2024-11-19	4.87	3.20	29.90	1.0
Lake Munson	2	78	2014-1-16	2024-11-21	18.55	6.39	171.68	1.0
Lake Bradford	1	43	2014-1-22	2024-11-26	5.51	3.40	34.18	1.0

The Leon County lakes are data scarce. Although the dataset contains 1,012 station-date observations, these observations are distributed across 11 lakes rather than representing repeated measurements from a single lake. Each lake has only 1–6 sampling locations and is monitored primarily on a quarterly basis. Consequently, only 43–252 observations are available for an individual lake over the 2014–2024 study period. The four lakes selected for spatial evaluation contain 252 observations for Lake Iamonia, 123 for Lake Jackson, 78 for Lake Munson, and only 43 for Lake Bradford. Thus, the lake-specific datasets used in this study are one to several orders of magnitude smaller than the large regional, national, and international datasets listed in Table 1. This difference in data structure is important for evaluating ML methods. Large datasets containing thousands to millions of observations can provide broad representation of spatial conditions, temporal variability, and different water quality states. In contrast, quarterly observations from a small number of locations provide substantially less information about short-term temporal dynamics and rare events of high Chl-a concentrations. Therefore, the objective of this study is not to demonstrate ML performance under a data-rich monitoring environment, but to evaluate whether commonly used ML methods can provide useful Chl-a simulations when being applied to the small datasets typically available from routine lake monitoring programs.

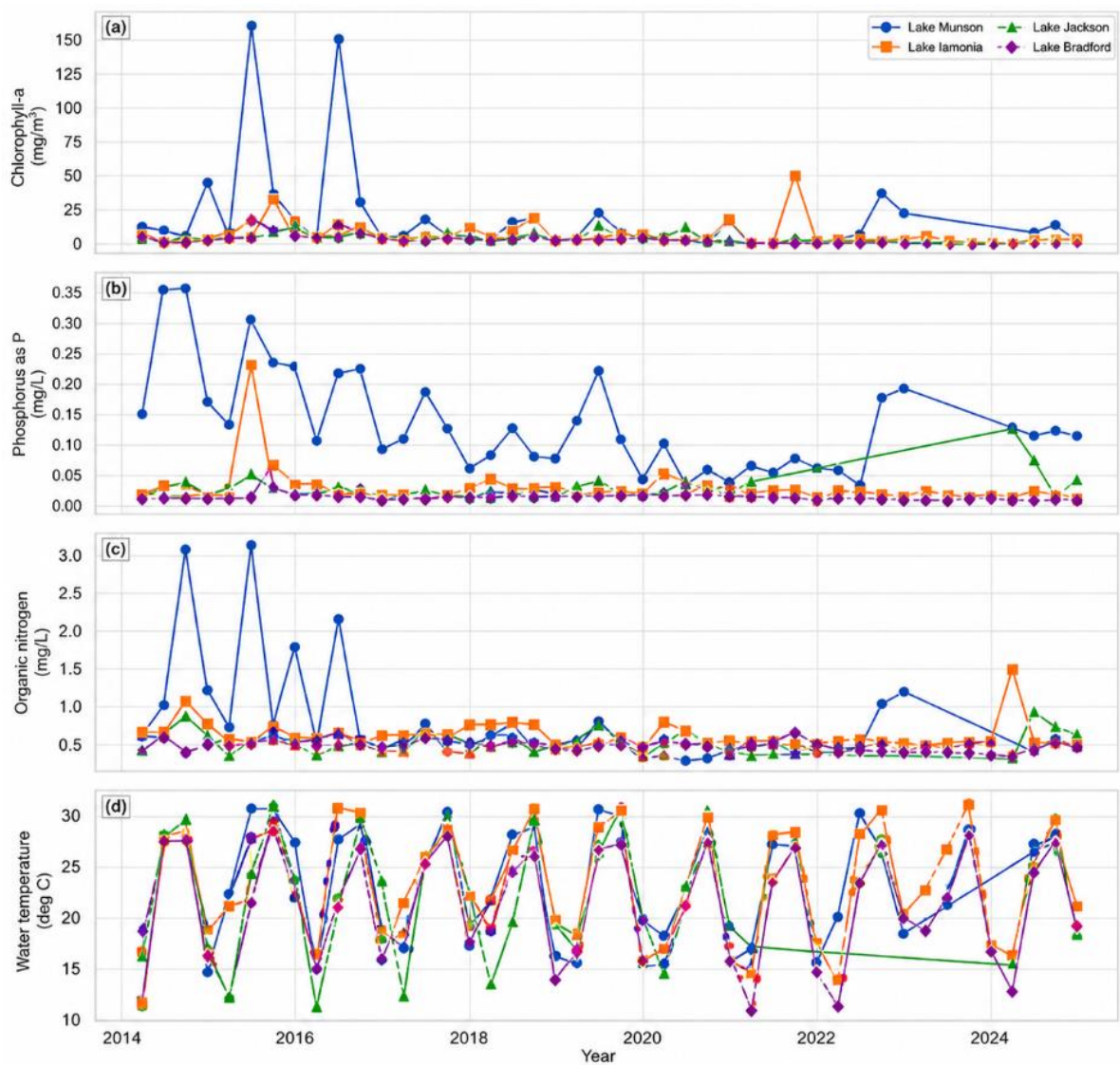


Figure 3. Time series of four water quality variables: (a) Chl-a, (b) total phosphorus, (c) organic nitrogen, and (d) water temperature for Lakes Iamonia, Jackson, Munson, and Bradford.

Figure 3 plots the time series of Chl-a, total phosphorus, organic nitrogen, and water temperature for the four lakes discussed above. Spatial variation is exhibited between the lakes. Lakes Iamonia and Munson have larger concentration values than Lakes Jackson and Bradford. This is not surprising, because Lake Jackson is an aquatic preserve and Lake Bradford is used for recreation. Temporal variation is also observed for each lake, especially for Lakes Iamonia and Munson. While Lake Munson has the highest Chl-a concentrations among the four lakes (Table 2), it has a small number of monitoring datasets (78 as listed in Table 2). This combination of data scarcity, temporal variation of water quality, and elevated Chl-a concentrations help explain why it is difficult to simulate Chl-a concentrations of Lake Munson, as discussed later.

Table 3. Basic statistics of selected water quality variables for all 11 lakes.

Variable	N	Mean	Median	Std. Dev.	Min	Max	Skewness
Chlorophyll-a (mg m-3)	1012	8.07	3.70	15.04	1.00	221.00	7.56
Phosphorus as P (mg/L)	1010	0.04	0.03	0.05	0.00	0.37	2.95
Organic Nitrogen (mg/L)	1007	0.60	0.54	0.49	0.09	13.30	17.81

Dissolved oxygen saturation (%)	1011	76.56	81.60	36.59	1.10	203.30	-0.12
pH	1011	6.57	6.53	0.99	4.12	9.97	0.46
Temperature (°C)	1011	22.88	23.09	6.04	6.96	34.01	-0.16
Turbidity (NTU)	1010	3.22	1.70	4.13	0.05	44.00	3.93
Total Suspended Solids (mg/L)	1009	5.33	5.00	3.47	0.57	50.00	5.62
Magnesium (mg/L)	910	4.62	3.56	3.81	0.00	36.70	1.37

Table 3 lists summary statistics of Chl-a and eight selected water quality variables. Chl-a was strongly right-skewed (skewness = 7.56), and several predictors were even more skewed (e.g., skewness = 17.81 for organic nitrogen). To correct this, each of the 43 predictors was Box-Cox transformed [35] with the power parameter, λ , estimated per variable to minimize skewness. Figure 4 compares the histogram of raw data and Box-Cox-transformed data for several predictor, and shows that the data skewness has been substantially reduced. The transformed data were subsequently standardized by using the z-score transformation using training-set means and standard deviations. Chl-a concentrations were log10-transformed to reduce the influence of high concentrations during training, and simulated Chl-a concentrations were inverse-transformed back to the original unit before computing root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) for cross-validation tests.

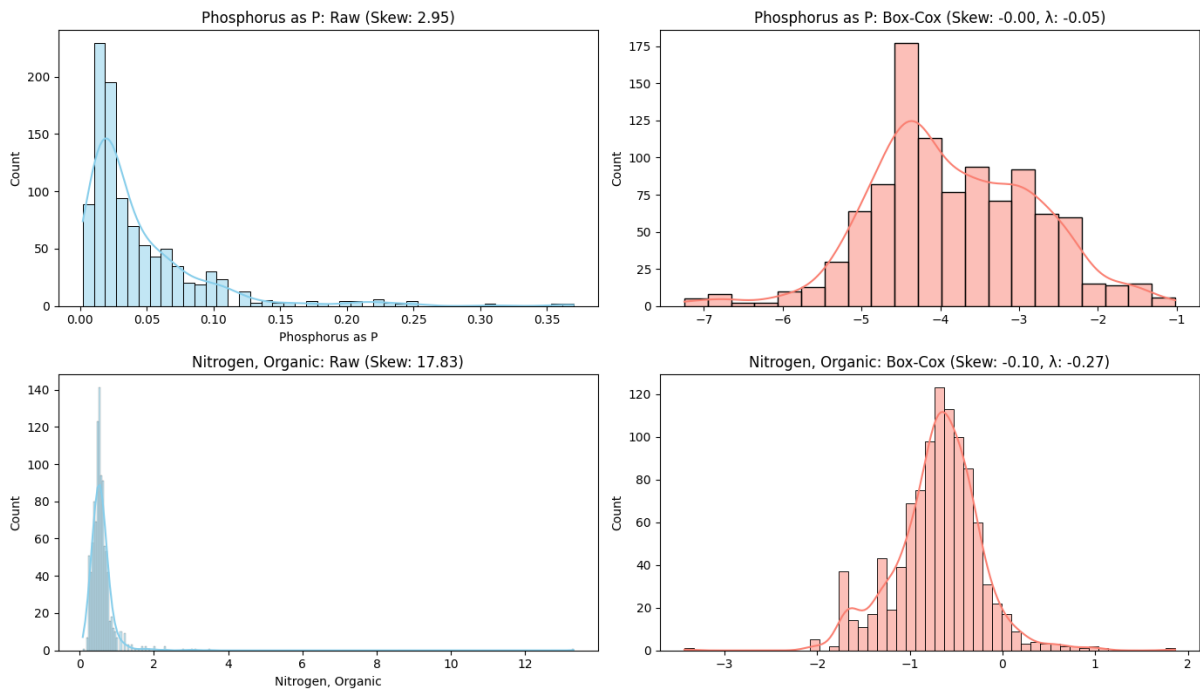


Figure 4. Histograms of selected water quality variables before and after the Box-Cox transformation used for data preprocessing. Values of λ estimated for the Box-Cox transformation are also shown in the figure. .

2.2. Machine Learning Methods

The Random Forest, XGBoost, and LSTM methods considered in this study represent three different assumptions about the data. Random Forest provides a robust baseline for nonlinear tabular data. It first trains many decision trees (each on a random subset of the data and predictors), and then averages their predictions – a strategy known as bagging – which makes it relatively insensitive to monotonic transformations and noisy predictors [21]. XGBoost uses a different strategy called regularized boosting, in which trees are added one at a time, each one correcting the errors left by the trees. Two safeguards keep this process from overfitting. One is shrinkage, which scales down each new tree’s contribution. The other is subsampling, which trains each tree on a random subset of the

training data. This approach often performs strongly on water-quality and environmental tabular data [19,22–24]. Evaluation of the two methods has been reported in literature, and it has been found that boosting methods frequently outperform individual decision-tree and neural-network approaches on nonlinear environmental datasets [15,18,20]. LSTM was chosen by this study as a deep-learning method, because recurrent networks are widely used in hydrologic and environmental studies when true temporal sequences are available [25–29].

In this study, Random Forest regression used bootstrap aggregation, 500 trees, square-root feature subsampling, and a minimum leaf size of three observations. XGBoost regression used 500 boosting rounds, a learning rate of 0.03, a maximum tree depth of three, subsampling, column subsampling, and squared-error loss. The LSTM was configured as a one-step neural-network benchmark using the standardized predictor vector for each station-date record, with one 32-unit LSTM layer, a dropout rate of 0.20, and a dense output layer. Because each input contained only one station-date vector rather than a multi-step sequence, this configuration could not meaningfully exploit long-term temporal memory. The restricted design was necessary because the monitoring record does not provide evenly spaced, station-specific sequences needed to support a recurrent time-series architecture.

2.3. Cross-Validation Tests for ML Method Evaluation

Three ML methods were evaluated using three cross-validation tests: five-fold cross-validation, chronological cross-validation, and spatial cross-validation. The five-fold cross-validation used five shuffled folds to estimate general interpolation skills for the entire dataset. The chronological cross-validation used the data of 2014 – 2019 for training and the data of 2020 – 2024 for testing. The spatial cross-validation held out a selected lake for testing and used the data of the remaining lakes for training. In the five-fold cross-validation, the full set of 1,012 station-date records was randomly shuffled and split into five equally sized folds; each fold served once as the test set while the remaining four folds were used for training, and preprocessing (imputation, Box-Cox transformation, and standardization) was refit on each training split to avoid information leakage. This test estimated how well a method interpolates within the overall dataset when training and testing data were drawn from the same lakes and time periods. As shown in Figure 2, data preprocessing was applied to training data to avoid information leakage, and the cross-validation tests and evaluation metrics were conducted for each training split and all three ML methods, so that any difference in ML simulation performance reflected method strength and weakness rather than inconsistent data handling.

For the five-fold cross-validation, each method was trained and tested five times (once per fold). The mean and standard deviation of RMSE, MAE, and R^2 across these five runs were then calculated and listed in Table 4. The standard deviation reflects how much performance varied from fold to fold, due to differences in which lakes and time periods fell into each test fold. The 95% confidence intervals for R^2 (shown in Figure 5) were calculated using paired bootstrap resampling. The out-of-fold prediction–observation pairs were resampled with replacement 1,000 times, R^2 was recalculated each time, and the 2.5th and 97.5th percentiles of these 1,000 values defined the performance interval.

The chronological and spatial cross-validation tests are structurally different from the five-fold test. Rather than randomly mixing observations, each test withholds an entire block of the data (a later time period or an entire lake) to evaluate a method’s ability to generalize rather than interpolate.

2.4. Feature Importance

Feature importance was evaluated using temporal-holdout permutation importance. Feature importance is evaluated only on the chronological holdout split, not on the five-fold or spatial splits, because the chronological split isolates the method’s ability to generalize forward in time on data it has never seen, which is the setting most relevant to operational forecasting. Each of the three ML methods was trained on the 2014 – 2019 data (the same training set used for the chronological cross-validation) and evaluated on the 2020 – 2024 data. This method-agnostic approach was chosen over method-specific metrics (e.g., Gini impurity for trees), because it allows a direct, standardized

comparison across the three fundamentally different methods. The temporal holdout technique also ensures that it reflects each feature’s true ability to simulate future observations without risking data leakage. For each predictor, the values of the predictor in the test set were randomly shuffled while all other predictors were held at their observed values, and the method’s RMSE was recalculated on this permuted test set. The increase in RMSE relative to the unpermuted baseline reflects the predictor’s importance, in that a large increase means the method relied heavily on the predictor to make accurate predictions. A little or no increase means the method’s simulation was insensitive to the predictor. This procedure was repeated multiple times for each predictor, and the average RMSE increases was used for identifying feature importance. Because importance is measured on held-out future data rather than on the training data or from method-internal splitting rules, it reflects each predictor’s contribution to genuine simulation skill rather than to in-sample fit, and it can be compared consistently across the three ML methods regardless of their internal structures.

3. Results

3.1. Five-Fold Cross-Validation

Table 4 reports five-fold cross-validation performance for the three methods. XGBoost had the best average performance (RMSE = 10.55 mg m⁻³, MAE = 4.14 mg m⁻³, R² = 0.476), followed by Random Forest and then LSTM. The fold-to-fold standard deviations were large for all methods, reflecting temporal variation in Chl-a within lakes and spatial variation across lakes. Figures 5a and 5d plot the average RMSE and R² values with their 95% bootstrap confidence intervals. The overlapping confidence intervals in Figure 5d indicate that, despite XGBoost’s numerical advantage, the three ML methods are not sharply distinguishable once resampling uncertainty is considered.

Table 4. Standardized five-fold cross-validation performance for all three methods.

Method	RMSE (mean)	RMSE (SD)	MAE (mean)	MAE (SD)	R ² (mean)	R ² (SD)
Random Forest	11.46	3.25	4.35	1.07	0.392	0.213
XGBoost	10.55	3.53	4.14	0.97	0.476	0.256
LSTM	11.74	3.48	4.54	1.15	0.373	0.199

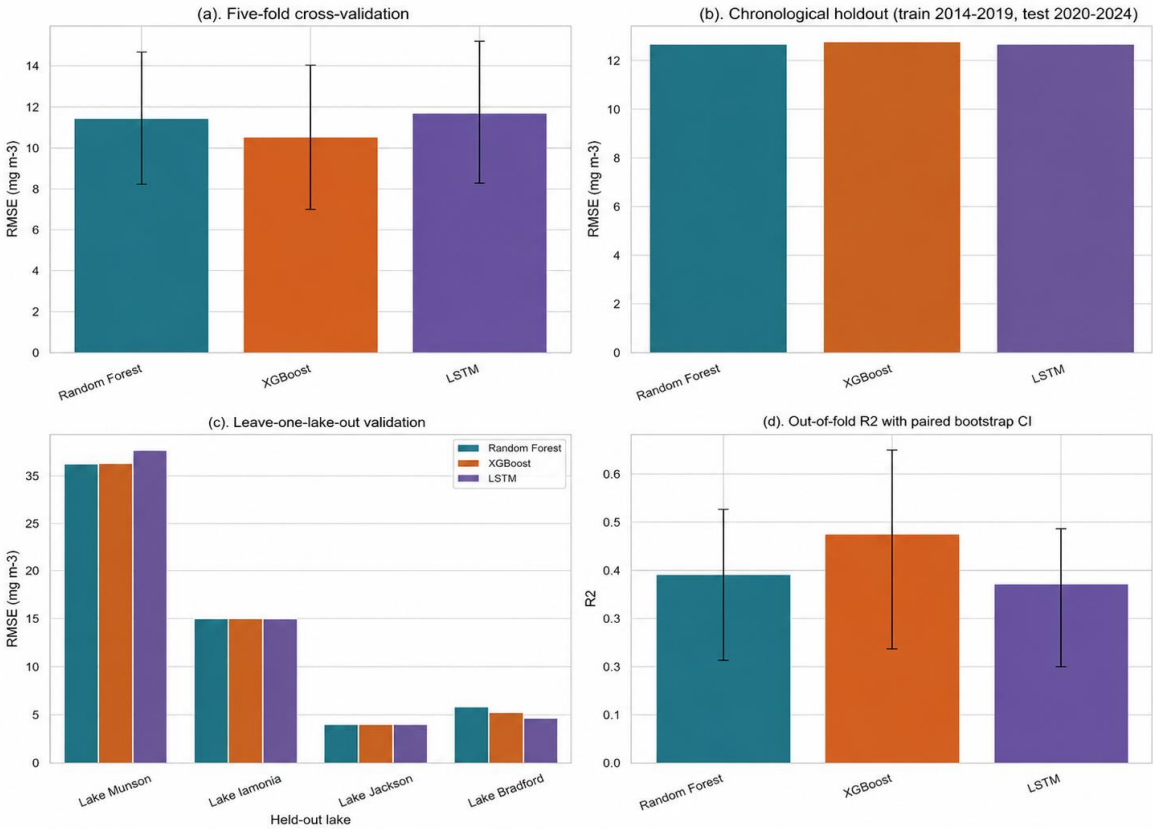


Figure 5. Statistical metrics of method performance for Random Forest, XGBoost, and LSTM across four validation approaches: (a) five-fold cross-validation RMSE (mg-m-3); (b) chronological holdout RMSE (trained 2014–2019, evaluated 2020–2024); (c) leave-one-lake-out spatial validation RMSE; and (d) five-fold cross-validation R2 with 95% confidence intervals derived from paired bootstrap resampling. All methods used identical preprocessing, predictors, and configurations.

3.2. Chronological and Spatial Cross-Validation

Table 6 lists the results of the chronological cross-validation. The three ML methods have similar performance, in that their RMSE, MAE, and R² values are similar, which is also shown in Figure 5b. The small difference in the statistical metrics between the methods indicates that method choice was less important than temporal distribution shift and bloom-event rarity in the future-period test. Comparing Figures 5a and 5b indicates that the five-fold cross-validation evaluation does not reveal simulation performance of future Chl-a concentrations.

Table 6. Chronological holdout performance after training on 2014-2019 observations and testing on 2020-2024 observations.

Method	Training Data	Testing Data	RMSE	MAE	R ²
Random Forest	631	381	12.69	4.27	0.063
XGBoost	631	381	12.83	4.37	0.043
LSTM	631	381	12.70	4.30	0.063

To better illustrate the statistical metrics listed in Tables 3 and 4, Figure 6 plots the measured and simulated Chl-a concentrations of the three ML methods for the five-fold cross-validation test (Figures 6a–c) and the chronological cross-validation test (Figures 6d–f). The RMSE, MAE, and R²

values shown in Figures 6a–c were computed on the out-of-fold simulations from all the five folds combined, and therefore differ slightly from the averaged values listed in Table 6. The figure shows that, while all the ML methods can reproduce low-to-moderate Chl-a concentrations, they consistently underestimated high Chl-a concentrations due to lack of such data.

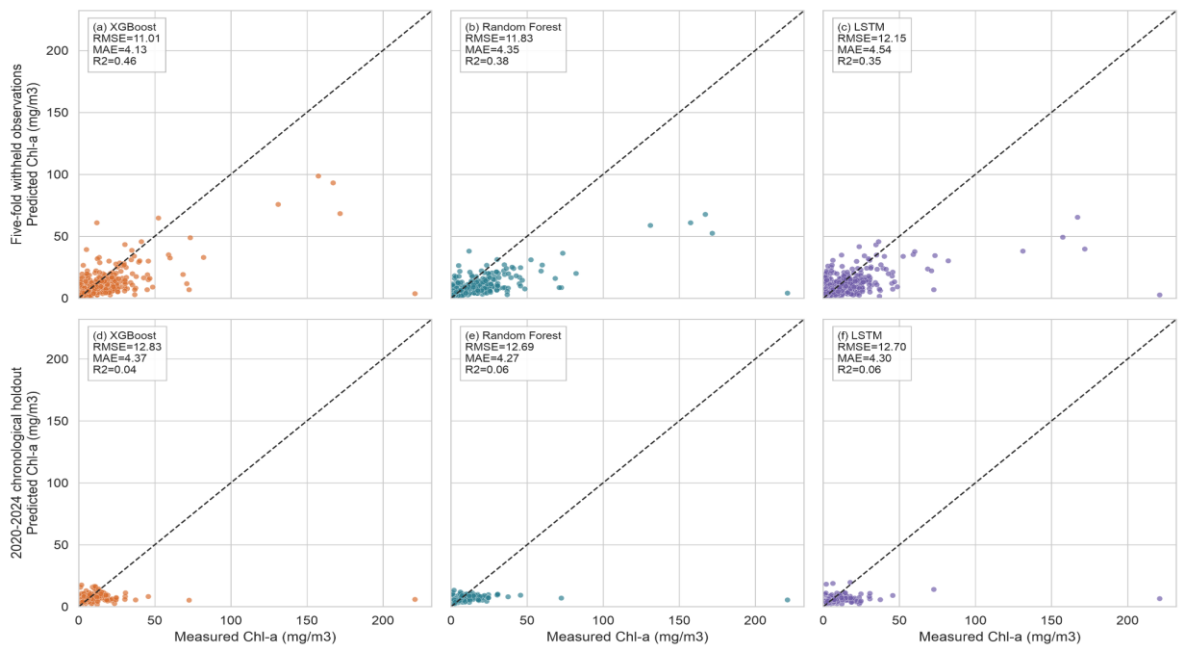


Figure 6. Measured and simulated Chl-a concentrations for (a–c) five-fold cross-validation and (d–f) chronological cross-validation.

Table 7 lists the results of the leave-one-lake-out spatial validation, and the RMSE values are plotted in Figure 5c. The table shows that simulation performance of the three ML methods varied sharply depending on which lake was held out. It indicates that simulation performance a method trained using 10 lakes does not transfer equally well to the held-out lake. Table 7 and Figure 5c indicate that, for each case of spatial cross-validation, the three ML methods have similar performance. The performance is the worst for Lake Munson and the best for Lake Jackson. RMSE (31.2 – 32.7 mg m⁻³) for Lake Munson was roughly eight times higher than that (4.0 – 4.1 mg m⁻³) for Lake Jackson. Similar differences were also observed for MAE and R². This is not surprising due to high Chl-a concentrations at Lake Munson (Table 6 and Figure 3). For the same reason, the performance is the second worst for Lake Iamonia. The performance is substantially better for Lake Jackson and Lake Bradford, due to relatively smaller variation of Chl-a concentrations at the two lakes. These results show that, for a lake with high Ch-a concentrations and large temporal variation, its Chl-a concentrations cannot be simulated by using ML methods trained using the data of the other lakes in the same region.

Table 7. Results of leave-one-lake-out spatial cross-validation for the four selected lakes.

Held-out lake	Method	Training Data	Testing Data	RMSE	MAE	R ²
Lake Munson	Random Forest	934	78	31.24	13.20	0.187
	XGBoost	934	78	31.35	13.36	0.181
	LSTM	934	78	32.69	13.21	0.110
Lake Iamonia	Random Forest	760	252	15.07	3.53	0.021
	XGBoost	760	252	15.07	3.38	0.020
	LSTM	760	252	14.99	3.52	0.030
Lake Jackson	Random Forest	889	123	4.06	2.78	0.280
	XGBoost	889	123	4.01	2.66	0.298
	LSTM	889	123	4.04	2.64	0.286
Lake Bradford	Random Forest	969	43	5.82	3.29	0.152
	XGBoost	969	43	5.25	3.42	0.311
	LSTM	969	43	4.62	3.16	0.465

3.3. Time-Series Simulation and Lake-Specific Predictions

To complement the cross-validation tests presented above, each method was fitted to all the data for the period of 2014 – 2024 to reconstruct the full Chl-a time series for each selected lake. These simulation results are presented for interpretive and visual purposes rather than as an additional cross-validation test. Figure 7 shows the observed and simulated Chl-a concentrations for all the three ML methods. The XGBoost simulations and observations (Figure 7a) clustered tightly along the 1:1 line across the full range of observed concentrations, including those with high Chl-a concentrations. Random Forest (Figure 7b) reproduced low-to-moderate Chl-a concentrations but could not simulate high Chl-a concentrations. In addition, the simulations deviated more from the observations, with the R² values decreasing from 0.93 for XGBoost from 0.85 for Random Forest. LSTM (Figure 7c) had the worst simulation performance, because LSTM had the most pronounced underestimation of the largest observed Chl-a concentration of 221 mg m⁻³. This is not surprising because LSTM is best suited to learning from long time series data rather than the intermittent data used in this study.

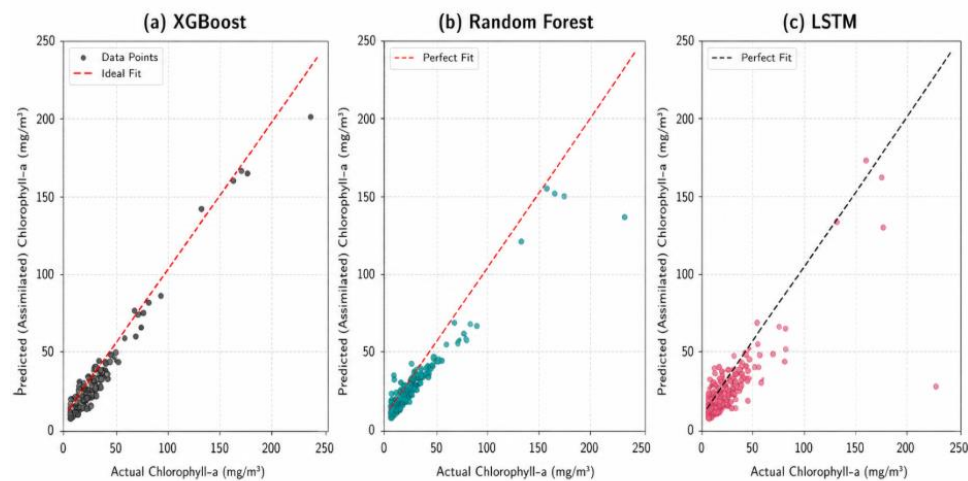


Figure 7. Simulated and observed Chl-a concentrations for (a) XGBoost, (b) Random Forest, and (c) LSTM methods when using the entire data for the study period of 2014 – 2024.

Figure 8 plots the observed and simulated Chl-a concentrations of the three ML methods for the four selected lakes. The figure illustrates how closely each method can track the timing, shape, and magnitude of observed Chl-a concentrations when using all the data, which is helpful for identifying which high Chl-a concentrations are difficult to reproduce. While all the three methods tracked the general temporal pattern in the four lakes, Random Forest and XGBoost underestimated high Chl-a concentrations for Lake Munson in 2015 and 2016 and for Lake Iamonia in 2021. Considering Table 5 and Figure 8 together indicates that, although the three ML methods are theoretically capable of simulating high Chl-a concentrations, it is necessary to have the high concentrations in the training data. Taking Lake Munson as an example, the three ML methods trained using the data of the other 10 lakes cannot simulate the high Chl-a concentrations of Lake Munson. It however is not the case for Lakes Jackson and Bradford, because the two lakes have low-to-moderate Chl-a concentrations, which is true for all the three ML methods.

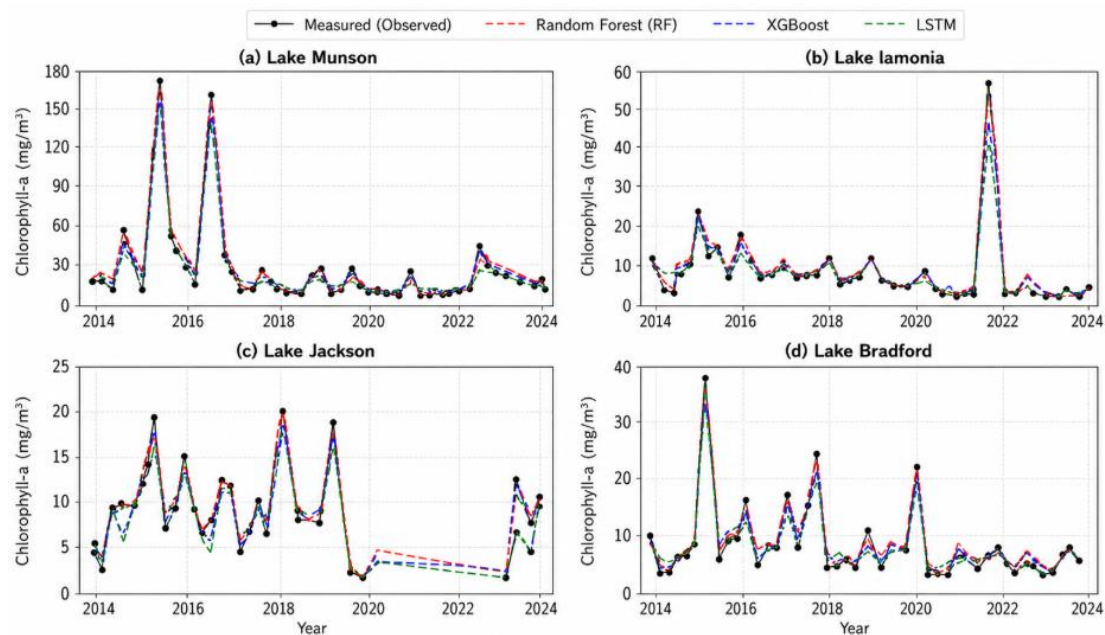


Figure 8. Simulated and observed time-series of Chl-a concentrations for (a) Lake Munson, (b) Lake Iamonia, (c) Lake Jackson, and (d) Lake Bradford.

3.4. Feature Importance

Figure 9 plots the identified top predictors, and shows that, across the difference importance rankings, Phosphorus as P was the top feature for Random Forest and XGBoost and remained among the top LSTM features. This is confirmed by Figure 10 that shows the Pearson correlation matrix among Chl-a and six of the water-quality variables most frequently identified as important across the three ML methods. Phosphorus as P has the largest correlation with Chl-a with the correlation coefficient $r = 0.43$. The relatively modest correlation values (with the maximum value of 0.43) implies that no single variable dominates the Chl-a simulations.

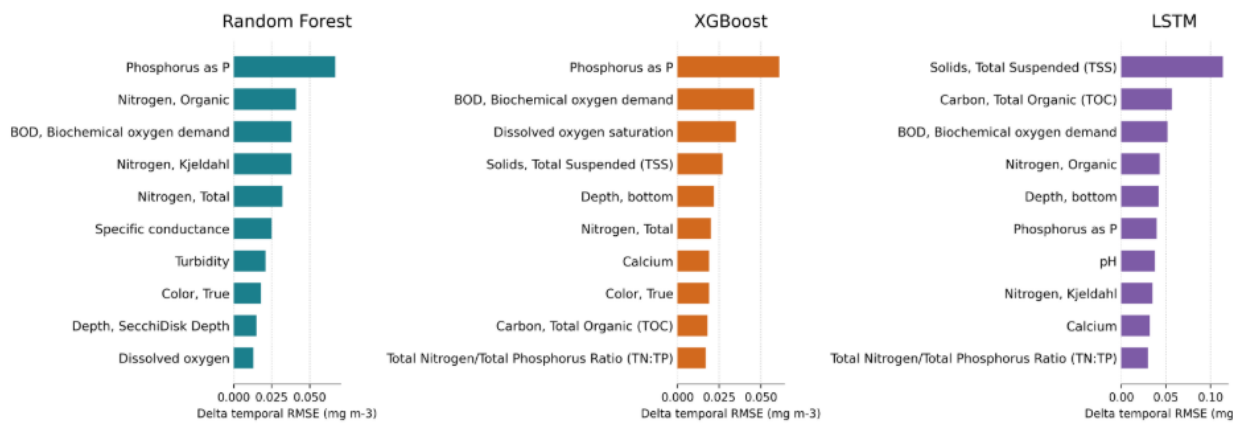


Figure 9. Temporal-holdout permutation importance for the three methods. Values show the increase in RMSE when a predictor was permuted in the 2020-2024 test set.

Figure 9 shows that, while nutrient- and oxygen-related predictors were identified as important across the methods, the exact rankings differed substantially. Random Forest also emphasized organic nitrogen, biochemical oxygen demand, total Kjeldahl nitrogen, and total nitrogen. XGBoost emphasized biochemical oxygen demand, dissolved oxygen saturation, total suspended solids, bottom depth, total nitrogen, calcium, and color. LSTM gave the largest importance to total suspended solids, followed by total organic carbon, biochemical oxygen demand, organic nitrogen, bottom depth, phosphorus, and pH. In particular, the LSTM importance profile differed substantially from those of Random Forest and XGBoost. Because the LSTM received only one station-date vector at a time, these differences cannot be attributed to learned long-term temporal dynamics. Instead, they reflected how the one-step neural-network benchmark weighted contemporaneous predictors under the sparse and irregular sampling design.

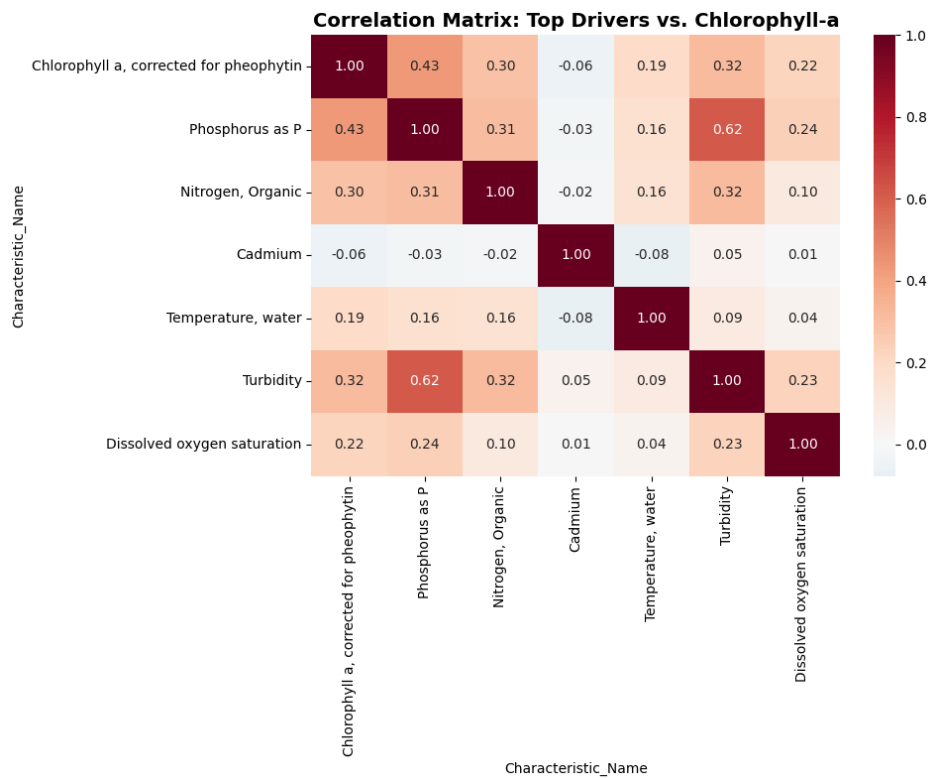


Figure 10. Pearson correlation matrix among Chl-a and six of the water-quality predictors most frequently identified as important by the temporal-holdout permutation analysis.

The important predictors identified by the ML methods are environmentally meaningful. Phosphorus and nitrogen were consistently ranked as important predictors, because they represent nutrient availability, which is a direct control on algal biomass in freshwater systems [1–6]. Biochemical oxygen demand and total organic carbon may indicate organic loading and microbial oxygen demand associated with productive or runoff-influenced conditions. Turbidity and total suspended solids can reflect runoff, sediment resuspension, and light limitation, while dissolved oxygen saturation and pH may respond to algal photosynthesis and respiration. Because some predictors are simultaneous observations rather than antecedent drivers, they should be interpreted as indicators associated with Chl-a conditions rather than necessarily causal inputs.

4. Discussion

For the five-fold cross-validation, XGBoost provided the best results measured by RMSE, MAE, and R², consistent with previous studies showing that ensemble tree methods often outperform deep-learning methods on tabular environmental datasets with relatively small sample sizes [19–24]. In contrast, LSTM generally provides its greatest advantages when long, regularly spaced temporal sequences are available for learning complex temporal dependencies [16,25–31]. Because the Leon County monitoring data consist primarily of intermittent quarterly sampling rather than continuous sensor measurements, the tree-based methods were better aligned with the structure of the available data. Random Forest remains competitive and may be preferable when interpretability, robustness, and ease of implementation are priorities.

The results of this study extend the data contrast identified in Table 1. Chen et al. [19], Zhao et al. [36], Pahlevan et al. [37], and Spaulding et al. [38] drew on regional, global, remote-sensing, or national-scale data comprising thousands to more than one million observations. In the present study, the lower-frequency, lake-specific record did not support a clear advantage for the more

complex LSTM model and yielded weak chronological simulations, indicating that model performance in large data-rich applications should not be assumed to carry over to data-scarce lakes.

The chronological and spatial cross-validation results are the central caution. For the chronological cross-validation, all the three methods had R^2 values near zero for 2020-2024, even though the R^2 values of the five-fold cross-validation were moderate. This raises a concern about prediction capability of the three methods. For the spatial cross-validation, the simulation results are substantially better for Lakes Jackson and Bradford than for Lakes Iamonia and Munson. These results suggest that, due to temporal and spatial variability of water quality among the lakes, it should be cautious when using the ML methods to simulate Chl-a concentrations for a lake with water quality conditions different from its previous conditions or from other lakes in the same region.

When the three ML methods were trained using all the data, Figure 8 shows that all three methods were capable of closely tracking the general temporal variation of Chl-a across the four selected lakes. The two largest Chl-a concentrations in 2015 and 2016 for Lake Munson were reconstructed well by every method. This confirms that the methods can simulate temporal variation of Chl-a concentrations for the lakes. However, for the high Chl-a concentration of Lake Iamonia in 2021, it was underestimated by all the three methods, especially LSTM. This is a common problem for ML methods, and is difficult to resolve until more monitoring data become available. It should be noted that the good match between the observed and simulated Chl-a concentrations in Figure 7 should be interpreted with caution. While it indicates that a method's ability to reconstruct historical data, it is not evidence of its ability to forecast Chl-a concentrations for a lake with substantially different conditions in space and/or time.

5. Conclusions

This study was focused on evaluating three ML methods (Random Forest, XGBoost, and LSTM) for simulating Chl-a concentrations in 11 data-scarce lakes located in Leon County, FL. Each lake has only 1 – 6 sampling locations with a quarterly monitoring schedule, and there are a total of 1,012 sets of observations (43 to 252 per lake) for the study period of 2014 – 2024. Each observation set includes 43 water quality variables, e.g., phosphorus, organic nitrogen, and dissolved oxygen. Box-Cox transformation and z-score standardization were applied to the water quality variables, and log-10 transformation was applied to the Chl-a concentrations. When all the data were used to train the three ML methods, each method satisfactorily reproduced the data, except for several high Chl-a concentrations. XGBoost had the best performance, which however was only slightly better than the performance of Random Forest and LSTM. This however does not mean that the three ML methods can be used to predict Chl-a concentration at a lake with substantially different water quality conditions from the data used to train the ML methods.

Simulation performance of the three ML methods was evaluated by using three cross-validation tests and the statistical metrics of RMSE, MAE, and R^2 . For the five-fold cross-validation, XGBoost had the best performance, while Random Forest and LSTM were slightly worse. For the chronological cross-validation, all the three ML methods had similar performance with R^2 near zero. For the spatial cross-validation, while Chl-a concentrations of Lake Jackson and Lake Bradford were satisfactorily simulated, the simulation for Lake Munson and Lake Iamonia was unsatisfactory due to high values and large temporal variations of Chl-a concentrations in the two lakes. These results led to the following conclusions: (1) for data-scarce lakes, sophisticated ML methods such as LSTM do not outperform relatively simpler methods such as XGboost and Random Forest; (2) when using trained ML methods to predict future Chl-a concentrations, the predictions should be used with caution, especially when the future lake conditions are different from the lake conditions of the training data; (3) when using trained ML methods to predict Chl-a concentrations of a new lake, the predictions are acceptable only when the lake conditions are similar to the lake conditions of the training data. Generally speaking, ML is still an effective tool for simulating Chl-a concentrations even for data-sparse lakes, whereas more monitoring data are needed for simulating high Chl-a concentrations with large temporal variations.

A method-agnostic importance test consistently identified a subset of water quality variables as top predictors across all the three ML methods, including phosphorus, nitrogen, biochemical oxygen demand, suspended solids, dissolved oxygen saturation, turbidity, pH, and related water-quality indicators. However, the feature importance ranking was different for different ML methods. The ranking of LSTM was not as reliable as those of XGBoost and Random Forest, because the training data were inadequate for fully train LSTM. The data scarcity is a limitation to this study. Future work should include hydrometeorological and watershed covariates, including rainfall, antecedent runoff, lake stage, watershed land cover, station coordinates, and seasonal terms.

6. Patents

Author Contributions: Conceptualization, A.W. and M.Y.; methodology, A.W.; software, A.W.; validation, A.W.; formal analysis, A.W.; investigation, A.W.; resources, M.Y.; data curation, A.W.; writing-original draft preparation, A.W.; writing-review and editing, M.Y.; visualization, A.W.; supervision, M.Y. All authors have read and agreed to the submitted version of the manuscript.

Funding: The second author was supported by US National Science Foundation (NSF) Collaborative Award Numbers 2536218 and 2536219.

Data Availability Statement: The water-quality data used in this study were obtained from the Leon County lake monitoring program. Analysis code is available online at <https://github.com/alexwang1500/LeonCountyCode/blob/main/munsonpaaper.py>.

Acknowledgments: The authors thank Johnny Richardson and Anna Padilla at Leon County, FL for providing lake water quality data.

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

References

1. Smith, V.H.; Schindler, D.W. Eutrophication science: where do we go from here? *Trends Ecol. Evol.* **2009**, *24*, 201-207.
2. Paerl, H.W.; Huisman, J. Climate change: a catalyst for global expansion of harmful cyanobacterial blooms. *Environ. Microbiol. Rep.* **2009**, *1*, 27-37.
3. Heisler, J.; Glibert, P.M.; Burkholder, J.M.; et al. Eutrophication and harmful algal blooms: A scientific consensus. *Harmful Algae* **2008**, *8*, 3-13.
4. Chorus, I.; Welker, M. *Toxic Cyanobacteria in Water: A Guide to Their Public Health Consequences, Monitoring and Management*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2021.
5. Carlson, R.E. A trophic state index for lakes. *Limnol. Oceanogr.* **1977**, *22*, 361-369.
6. OECD. *Eutrophication of Waters: Monitoring, Assessment and Control*; Organisation for Economic Co-operation and Development: Paris, France, 1982.
7. Gholizadeh, M.H.; Melesse, A.M.; Reddi, L. A comprehensive review on water quality parameters estimation using remote sensing techniques. *Sensors* **2016**, *16*, 1298.
8. Glasgow, H.B.; Burkholder, J.M.; Reed, R.E.; et al. Real-time remote monitoring of water quality: a review of current applications, and advancements in sensor, telemetry, and computing technologies. *J. Exp. Mar. Biol. Ecol.* **2004**, *300*, 409-448.
9. Reichstein, M.; Camps-Valls, G.; Stevens, B.; et al. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195-204.
10. Bergen, K.J.; Johnson, P.A.; de Hoop, M.V.; Beroza, G.C. Machine learning for data-driven discovery in solid Earth geoscience. *Science* **2019**, *363*, eaau0323.
11. Shen, C. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resour. Res.* **2018**, *54*, 8558-8593.

12. Maier, H.R.; Dandy, G.C. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environ. Model. Softw.* **2000**, *15*, 101-124.
13. Kim, Y.H.; Im, J.; Ha, H.K.; et al. Machine learning approaches to coastal water quality monitoring using GOCI satellite data. *GISci. Remote Sens.* **2014**, *51*, 158-174.
14. Hafeez, S.; Wong, M.S.; Ho, H.C.; et al. Comparison of machine learning algorithms for retrieval of water quality indicators in case-II waters: A case study of Hong Kong. *Remote Sens.* **2019**, *11*, 617.
15. Zhu, S.; Hrnjica, B.; Ptak, M.; et al. Forecasting of water level in multiple temperate lakes using machine learning methods. *J. Hydrol.* **2020**, *585*, 124819.
16. Barzegar, R.; Aalami, M.T.; Adamowski, J. Short-term water quality variable prediction using a hybrid CNN-LSTM deep learning model. *Stoch. Environ. Res. Risk Assess.* **2020**, *34*, 415-433.
17. Jiang, Y.; Zhang, L.; Zhang, B.; et al. Modeling daily chlorophyll-a dynamics in a German lowland river using artificial neural networks and multiple linear regression approaches. *Limnology* **2014**, *15*, 47-56.
18. Tiyasha; Tung, T.M.; Yaseen, Z.M. A survey on river water quality modelling using artificial intelligence methods: 2000-2020. *J. Hydrol.* **2020**, *585*, 124670.
19. Chen, K.; Chen, H.; Zhou, C.; et al. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning methods based on big data. *Water Res.* **2020**, *171*, 115454.
20. Heddiam, S.; Ptak, M.; Zhu, S. Modelling of daily lake surface water temperature from air temperature: Extremely randomized trees versus Air2Water, MARS, M5Tree, RF and MLPNN. *J. Hydrol.* **2020**, *588*, 125130.
21. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5-32.
22. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13-17 August 2016; pp. 785-794.
23. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **2021**, *54*, 1937-1967.
24. Huang, G.; Wu, L.; Ma, X.; et al. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. *J. Hydrol.* **2019**, *574*, 1029-1041.
25. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735-1780.
26. Kratzert, F.; Klotz, D.; Brenner, C.; et al. Rainfall-runoff modelling using Long Short-Term Memory networks. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 6005-6022.
27. Hu, C.; Wu, Q.; Li, H.; et al. Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. *Water* **2018**, *10*, 1543.
28. Zhou, J.; Wang, Y.; Xiao, F.; et al. Water quality prediction method based on IGRA and LSTM. *Water* **2018**, *10*, 1148.
29. Baek, S.S.; Pyo, J.; Chun, J.A. Prediction of water level and water quality using a CNN-LSTM combined deep learning approach. *Water* **2020**, *12*, 3399.
30. Park, Y.; Cho, K.H.; Park, J.; et al. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning methods in freshwater and estuarine reservoirs, Korea. *Sci. Total Environ.* **2015**, *502*, 31-41.
31. Zhi, W.; Feng, D.; Tsai, W.P.; et al. From hydrometeorology to river water quality: Can a deep learning model predict dissolved oxygen at the continental scale? *Environ. Sci. Technol.* **2021**, *55*, 2357-2368.
32. Hoyer, M.V.; Frazer, T.K.; Notestein, S.K.; Canfield, D.E. Vegetative characteristics of three low-lying Florida coastal rivers in relation to flow, light, salinity and nutrients. *Hydrobiologia* **2002**, *485*, 217-238.
33. Brenner, M.; Schelske, C.L.; Keenan, L.W. Historical rates of sediment and nutrient accumulation in marshes of the Upper St. Johns River Basin, Florida, USA. *J. Paleolimnol.* **2001**, *26*, 241-257.
34. Kindinger, J.L.; Davis, J.B.; Flocks, J.G. Geology and evolution of lakes in north-central Florida. *Environ. Geol.* **1999**, *38*, 301-321.
35. Box, G.E.P.; Cox, D.R. An analysis of transformations. *J. R. Stat. Soc. Ser. B Methodol.* **1964**, *26*, 211-252.

36. Zhao, L.; Zhu, R.; Zhou, Q.; Jeppesen, E.; Yang, K. Trophic status and lake depth play important roles in determining the nutrient-chlorophyll a relationship: Evidence from thousands of lakes globally. *Water Res.* **2023**, *235*, 120182.
37. Pahlevan, N.; Smith, B.; Schalles, J.; Binding, C.; Cao, Z.; Ma, R.; Alikas, K.; Kangro, K.; Gurlin, D.; Hà, N.; Matsushita, B.; Moses, W.; Greb, S.; Lehmann, M.K.; Ondrusek, M.; Oppelt, N.; Stumpf, R. Seamless retrievals of chlorophyll-a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach. *Remote Sens. Environ.* **2020**, *240*, 111604.
38. Spaulding, S.A.; Platt, L.R.C.; Murphy, J.C.; Covert, S.A.; Harvey, J.W. Chlorophyll a in lakes and streams of the United States (2005–2022). *Sci. Data* **2024**, *11*, 611.
39. Figueiredo, T.A.; Souza, B.T.A.; Salim, D.H.C.; Mello, C.C.S.; Pereira, G.; Amorim, C.C. Proximal hyperspectral sensing and machine learning for chlorophyll-a retrieval in optically complex urban freshwaters. *Limnol. Rev.* **2026**, *26*, 32.
40. Cubilla-Montilla, M.; Castillo, M.; Carrasco, G.; Torres-Cubilla, C.A. Key physicochemical and biological factors associated with chlorophyll-a concentrations: A machine learning approach. *Limnol. Rev.* **2026**, *26*, 46.

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.