



Article

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

Alexander Wang¹ and Ming Ye^{2,*} ¹ College of Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA; awang871@gatech.edu² Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, FL 32306, USA

* Correspondence: mye@fsu.edu

Abstract

Chlorophyll-a (Chl-a) is widely used as a proxy for phytoplankton biomass and lake trophic status. This exploratory study compared Random Forest, XGBoost, and a one-step Long Short-Term Memory (LSTM) network using 1012 station-date observations and 43 contemporaneous water quality predictors from Leon County, Florida, during 2014–2024. Initial shuffled five-fold validation favored XGBoost (mean RMSE 10.45 $\mu\text{g L}^{-1}$; mean R^2 0.484), but shared stations across folds and prior manual configuration selection limit that comparison. Additional validation held out entire stations in five outer folds and selected configurations using three station-blocked inner folds. Pooled outer RMSE was 11.92, 10.49, and 10.91 $\mu\text{g L}^{-1}$ for Random Forest, XGBoost, and LSTM, respectively; the corresponding R^2 values were 0.371, 0.513, and 0.473. Six expanding training periods ending in 2017–2022, each assessed over the next two years, evaluated sensitivity to chronological cutoff. These tests complement the original 2014–2019 versus 2020–2024 split, for which R^2 was 0.054–0.071. Random Forest had the lowest RMSE in all six windows, while LSTM had negative R^2 in four. Spatial transfer remained limited for Lakes Munson and Iamonia. Nested evaluation separates the new configuration search from each outer assessment, but the dataset and predictor panel had already been examined; the reanalysis is not an untouched external validation. The findings describe conditional concentration estimation using contemporaneous measurements, not advance forecasting or assured transfer across contrasting lakes.

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



Academic Editor: Mariusz Ptak

Received: 5 September 2026

Revised: 10 October 2026

Accepted: 16 October 2026

Published: 20 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

1. Introduction

Eutrophication is a persistent water quality concern because 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) is widely used as an indicator of algal biomass and trophic state [5,6]. Machine learning (ML) methods can represent nonlinear relationships among water quality variables and have been applied to Chl-a estimation and other water resource problems [7–11]. The size and structure of available datasets vary substantially among applications. Table 1 summarizes selected examples. Chen et al. [12] compared 10 learning models using 33,612 observations from major rivers and lakes in China during 2012–2018. Zhao et al. [13] examined nutrient–Chl-a relationships using 25,083 observations from 2849 lakes. Pahlevan et al. [14] developed an ML approach for Chl-a retrieval using field

and remote-sensing observations. Spaulding et al. [15] compiled more than 1.37 million pigment measurements from approximately 84,000 U.S. sites during 2005–2022; this resource includes multiple pigment types and both water-column and benthic samples, so its total is not equivalent to a single Chl-a modeling dataset. Remote sensing and automated monitoring can provide spatial coverage or sampling frequencies unavailable from conventional grab sampling [16–19]. However, large spatial coverage does not necessarily imply dense water quality observations at each site. For example, Zhi et al. [20] combined sparse dissolved oxygen measurements from 236 watersheds with daily hydrometeorological inputs.

Table 1. Dataset size and scope in selected water quality and Chl-a studies. Counts refer to the records or sites described by each study and are not directly equivalent measures of training-set size.

Reference	Dataset Size or Scope	Data Source	Study Application
Chen et al. (2020) [12]	33,612	Major rivers and lakes in China, 2012–2018	Surface water quality prediction
Zhao et al. (2023) [13]	25,083	Global dataset covering 2849 lakes worldwide	Nutrient–Chl-a relationships across lakes
Pahlevan et al. (2020) [14]	2943	Satellite/field matchup observations across diverse water bodies	Chl-a retrieval in inland and coastal waters
Zhi et al. (2021) [20]	236 watersheds	Sparse dissolved oxygen and daily hydrometeorological data across the United States	Continental-scale water-quality prediction
Spaulding et al. (2024) [15]	More than 1,374,000 pigment measurements	United States, 2005–2022	Harmonized pigment compilation for modeling applications
Figueiredo et al. (2026) [21]	82	Matched hyperspectral-Chl-a observations from Pampulha Reservoir, Brazil	Chl-a retrieval in an optically complex urban reservoir
Cubilla-Montilla et al. (2026) [22]	1176	Monthly observations from 14 stations in Lake Gatun, Panama, 2017–2023	Chl-a prediction and predictor interpretation

Many lake monitoring programs collect monthly or quarterly grab samples at a limited number of stations, producing small datasets with irregular intervals, missing predictors, and uneven representation among lakes. Site-specific applications reported in *Limnological Review* illustrate this range: Figueiredo et al. [21] retained 82 matched hyperspectral–Chl-a observations from a reservoir, whereas Cubilla-Montilla et al. [22] analyzed monthly observations from 14 stations in Lake Gatun during 2017–2023. These studies show that ML can be informative in individual water bodies, but performance depends on the sampling design, predictors, and validation strategy. A practical question is therefore how well commonly used ML methods estimate Chl-a in a sparse routine monitoring network, particularly when evaluated in later years or in a lake excluded from training.

This study examined this question using the Leon County, Florida, lake monitoring network (Figure 1). The analysis included 1012 station–date observations during 2014–2024. Four focal lakes selected for spatial evaluation had 1–6 sampling stations and 43–252 observations each, primarily from quarterly monitoring. The pooled dataset therefore contains observations from multiple stations, lakes, and years rather than 1012 independent observations of one lake. Sparse sampling limits representation of short-term changes and infrequent high Chl-a concentrations.

This study evaluated two ensemble tree methods, Random Forest and XGBoost, and a one-step Long Short-Term Memory (LSTM) network. Ensemble tree methods can capture nonlinear relationships and interactions in tabular environmental data [12,23–27]. LSTM networks are widely used for hydrologic and water quality sequences [28–32], but the present implementation received one contemporaneous predictor vector per station–date observation. It was therefore evaluated as a neural-network benchmark, not as a model of long-term temporal memory. The contribution of this study is the comparison of these

methods under sparse monitoring conditions and different validation settings, rather than the development of a new algorithm.

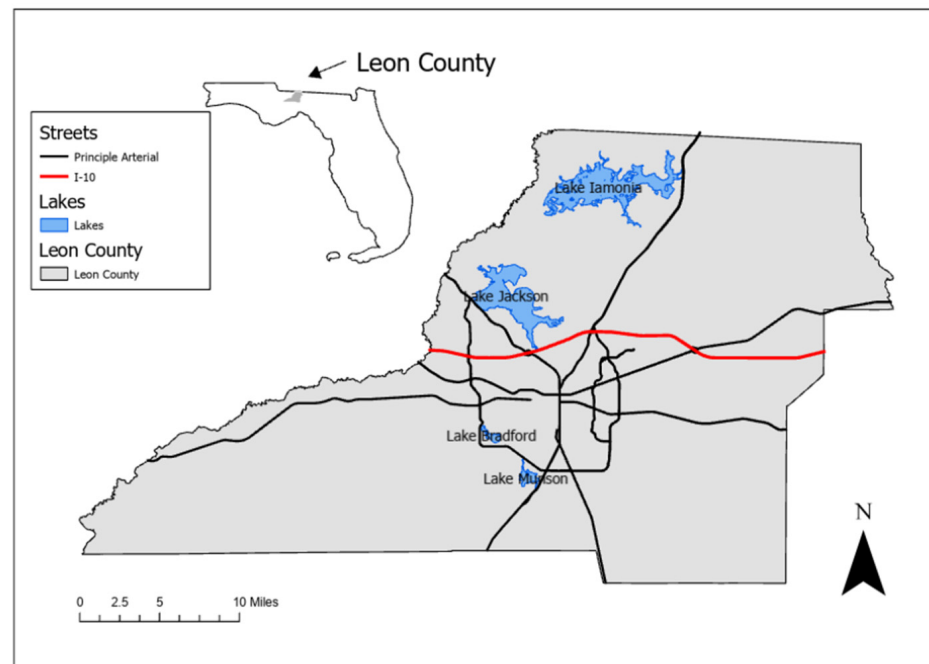


Figure 1. Location of Leon County, Florida, and the four lakes selected for spatial evaluation: Iamonia, Jackson, Munson, and Bradford.

Validation design affects the interpretation of model performance. Shuffled splits can place observations from the same station or sampling period in both training and test sets, yielding optimistic estimates of transfer to new conditions. Likewise, fitting a historical time series does not establish forecasting skill. The initial analysis used shuffled five-fold cross-validation, a 2014–2019 versus 2020–2024 chronological holdout, and four focal-lake holdouts. The shuffled results describe interpolation within the sampled network and retain within-station dependence. Because configurations were selected after exploratory comparisons, the initial estimates may also be optimistic. We therefore added five station-blocked outer folds with nested configuration selection and six expanding-window temporal holdouts. All models shared the observation splits, and preprocessing was fitted within each applicable training set. These additional assessments address station overlap and cutoff sensitivity while retaining the limitations of a retrospective reanalysis.

This study addressed three questions: (1) Which evaluated configuration has the best observed performance under exploratory shuffled five-fold cross-validation? (2) How much performance is retained when models trained on 2014–2019 observations are evaluated on 2020–2024 observations? (3) How well do models estimate Chl-a in a focal lake excluded from training? Additional tests examined whether comparisons persisted when entire stations were excluded, configurations were selected within training data, and chronological cutoffs varied. Temporal-holdout permutation importance was also used to identify predictors associated with model performance. Because the predictors were measured contemporaneously with Chl-a, the analyses assess concentration estimation conditional on available water quality measurements, not advance forecasts of future blooms.

2. Materials and Methods

2.1. Lake Water Quality Data and Preprocessing

Leon County, Florida, monitors lake water quality primarily quarterly, although sampling frequency varies among stations and years. Monitoring records from 2014–2024 were used in this study. Dates and numeric results were parsed, and records without valid dates, station names, characteristic names, or numeric values were excluded. Measurements were pivoted to a station–date matrix by taking the arithmetic mean of all numeric records sharing a station, date, and characteristic. The grouping did not distinguish sampling time, activity identifier, replicate, or depth and should therefore be described as aggregation of repeated records rather than removal of identical duplicates. Records without chlorophyll-*a* corrected for pheophytin were removed, and all chlorophyll and pheophytin variables were excluded from the predictors. Variables were retained when at least 100 of the 1012 response records had nonmissing values. This pooled availability screen produced 43 predictors, listed with their missingness rates in Table S1. Imputation and transformation parameters were subsequently fitted within each training split (Figure 2).

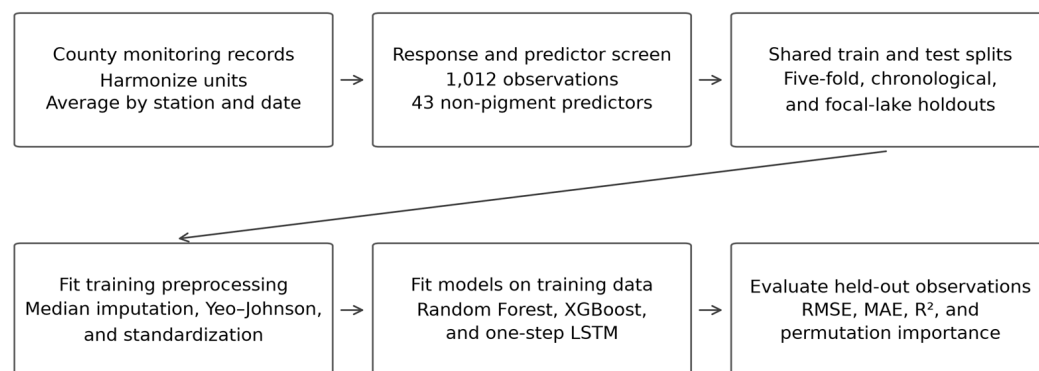


Figure 2. Data preparation and model evaluation workflow. Unit harmonization precedes aggregation into station–date records. Shared train–test splits precede fitting of imputation and transformation parameters. Each fitted preprocessing pipeline is applied unchanged to its test set. The initial evaluation designs are supplemented by the station-blocked and repeated temporal assessments in Section 2.5.

Before station–date aggregation, source concentration units were harmonized for boron, calcium, hardness (Ca + Mg), magnesium, and titanium. Values labeled in mg/L were multiplied by 1000 and expressed in $\mu\text{g/L}$; values already labeled in $\mu\text{g/L}$ were unchanged. Unit spelling and capitalization variants were treated equivalently. Chl-*a* was recorded in mg/m^3 , which is numerically equivalent to $\mu\text{g/L}$, and is reported here in $\mu\text{g/L}$. All three models were refitted, and all validation metrics, bootstrap intervals, and permutation importances were recalculated from the unit-harmonized dataset. The study period, response observations, predictor set, and validation splits were retained.

Across the 43 predictors, 4342 of 43,516 entries were missing (9.98%); variable-specific missingness ranged from 0.10% to 82.61% (Table S1). Training-set median imputation was used as a simple, stable baseline for sparse, skewed data. It does not reconstruct seasonal variation, and no comparison with seasonal or multiple imputation was performed. The high missingness of several microbial and trace-element variables limits interpretation of their effects.

Organic nitrogen denotes the county-reported “Nitrogen, Organic” variable in mg N L^{-1} . Its source method metadata identify calculation as total Kjeldahl nitrogen minus ammonia nitrogen ($\text{TKN} - \text{NH}_3\text{-N}$); it was used as supplied rather than recalculated for modeling. Total Kjeldahl nitrogen (TKN) was reported under EPA Method 351.2; ammonia records primarily used EPA Method 350.1, with two records under SM 4500-NH₃ G. Organic nitrogen represents nitrogen bound in organic material, including living biomass

and detritus, and is not synonymous with immediately bioavailable dissolved nitrogen. Total phosphorus refers to the source “Phosphorus as P” variable, separately recorded from “Phosphorus, orthophosphate as P.” These variables were retained as distinct predictors.

An audit of numeric source records contributing to the retained response and predictors found 2230 station–date–characteristic groups with multiple entries. Two groups had more than one recorded actual depth: Chl-a at station MI2 on 30 October 2018 and Secchi depth at H07 on 9 May 2019. All entries within each group were averaged consistently, including these cases. Sampling-depth and replicate distinctions were not preserved in the model matrix, so the aggregated records should not be interpreted as depth-specific measurements.

Lakes Iamonia, Jackson, Munson, and Bradford were selected for spatial evaluation to represent contrasting locations and water quality conditions (Figure 1). Table 2 summarizes their sampling coverage and Chl-a distributions. Lake Iamonia is connected intermittently to the Ochlockonee River and experiences variable water levels [33]. Lake Jackson is a sinkhole-influenced aquatic preserve with substantial hydrologic fluctuations [34]. Lake Munson receives water from developed areas south of Tallahassee and has a history of nutrient loading and restoration efforts [35]. These differences provide context for assessing transfer between lakes, without implying that land use or management designation alone determines Chl-a concentration. The four lakes are case studies rather than an exhaustive or randomly selected sample of all lakes. Their observed Chl-a means, maxima, sample counts, and station coverage in Table 2 describe their contrasts; no prespecified numerical selection rule or representativeness test was used.

Table 2. Sampling coverage and Chl-a statistics for the four focal lakes during 2014–2024. Concentrations are in $\mu\text{g L}^{-1}$; observation counts refer to station–date records.

Lake	Stations	Records	First Date	Last Date	Mean	Median	Max	Min
Lake Iamonia	6	252	20 February 2014	30 October 2024	5.66	2.90	221.00	1.0
Lake Jackson	4	123	19 February 2014	19 November 2024	4.87	3.20	29.90	1.0
Lake Munson	2	78	16 January 2014	21 November 2024	18.55	6.39	171.68	1.0
Lake Bradford	1	43	22 January 2014	26 November 2024	5.51	3.40	34.18	1.0

The four focal lakes contributed 252 observations for Iamonia, 123 for Jackson, 78 for Munson, and 43 for Bradford. Pooling records improves the number of observations available for training but does not create dense temporal coverage for any individual lake. Compared with the larger compilations in Table 1, these lake-specific records provide limited information about short-term dynamics and rare high-concentration events. The analysis consequently focuses on whether routine monitoring data support useful concentration estimates and on the limits of transferring those estimates to new times or locations.

Figure 3 shows temporal and spatial differences in Chl-a, total phosphorus, organic nitrogen, and water temperature. Lake Munson had the highest mean Chl-a concentration among the focal lakes ($18.55 \mu\text{g L}^{-1}$), whereas the highest individual observation occurred in Lake Iamonia ($221 \mu\text{g L}^{-1}$; Table 2). Same-date lake means in Figure 3 can be lower than individual station maxima in Table 2. Lake Munson contributed only 78 station–date observations despite its high concentrations and temporal variability. This uneven representation helps contextualize, but does not by itself establish the cause of, the prediction errors examined below.

Chl-a and several predictors were strongly right-skewed (Table 3). Predictors were median-imputed, transformed using the Yeo–Johnson power transformation [36], and standardized using parameters fitted to the training data. Unlike the Box–Cox transformation, Yeo–Johnson accommodates zero values present in the predictor matrix. The transformation parameter was estimated separately for each predictor by maximum likelihood. Figure 4 illustrates the transformation for two predictors using the chronological training period.

The response was transformed as $\ln(1 + \text{Chl-a})$ and standardized using its training-set mean and standard deviation. Predictions were returned to the original scale by reversing standardization and applying $\exp(z) - 1$, then clipped at zero. RMSE, MAE, and R^2 were calculated in the original concentration scale. No additional retransformation bias correction was applied.

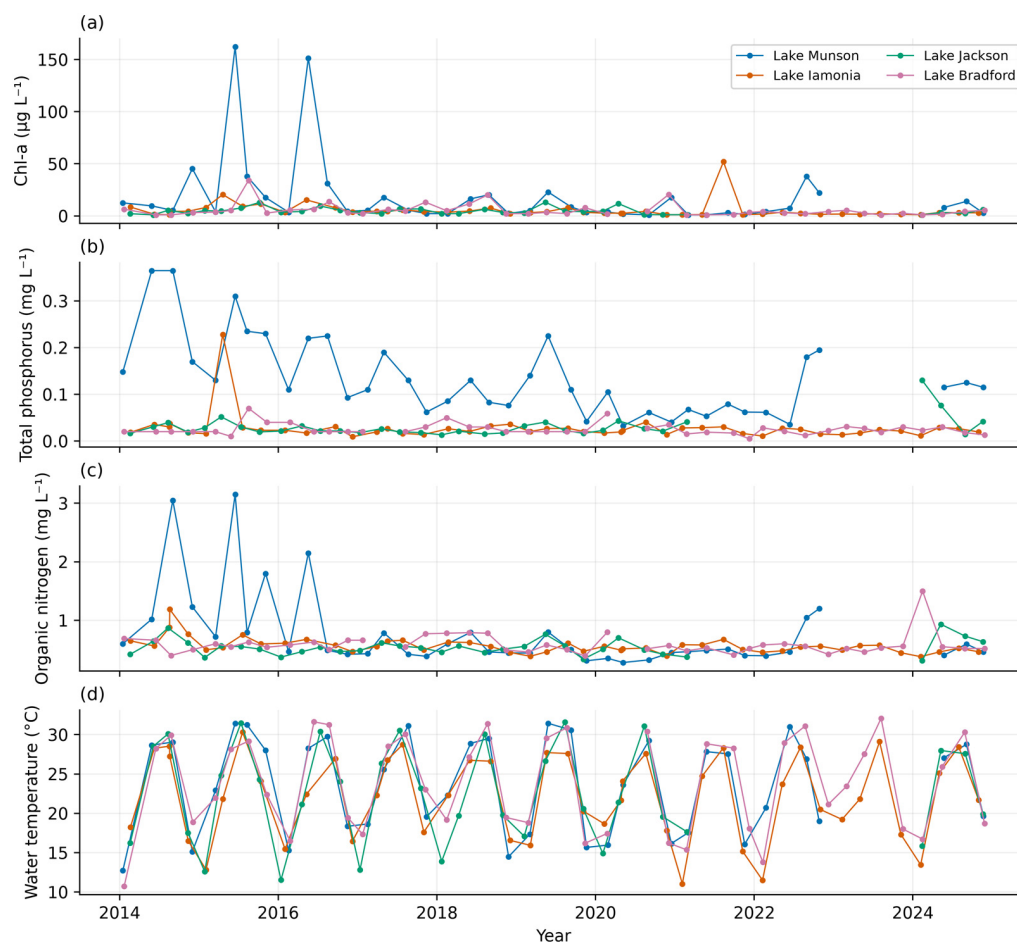


Figure 3. Observed water quality in the four focal lakes: (a) Chl-a, (b) total phosphorus, (c) organic nitrogen, and (d) water temperature. Points are means of available station measurements within each lake on the same date. Lines are visual guides and are broken when successive available observations are more than 180 days apart; no observations are interpolated.

Table 3. Summary statistics for Chl-a and selected predictors across the pooled dataset. N is the number of nonmissing observations before imputation.

Variable	N	Mean	Median	Std. Dev.	Min	Max	Skewness
Chlorophyll-a ($\mu\text{g/L}$)	1012	8.07	3.70	15.04	1.00	221.00	7.56
Total phosphorus (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 ($^{\circ}\text{C}$)	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 ($\mu\text{g/L}$)	910	5070.07	3705.00	3515.25	484.00	36,700.00	1.68

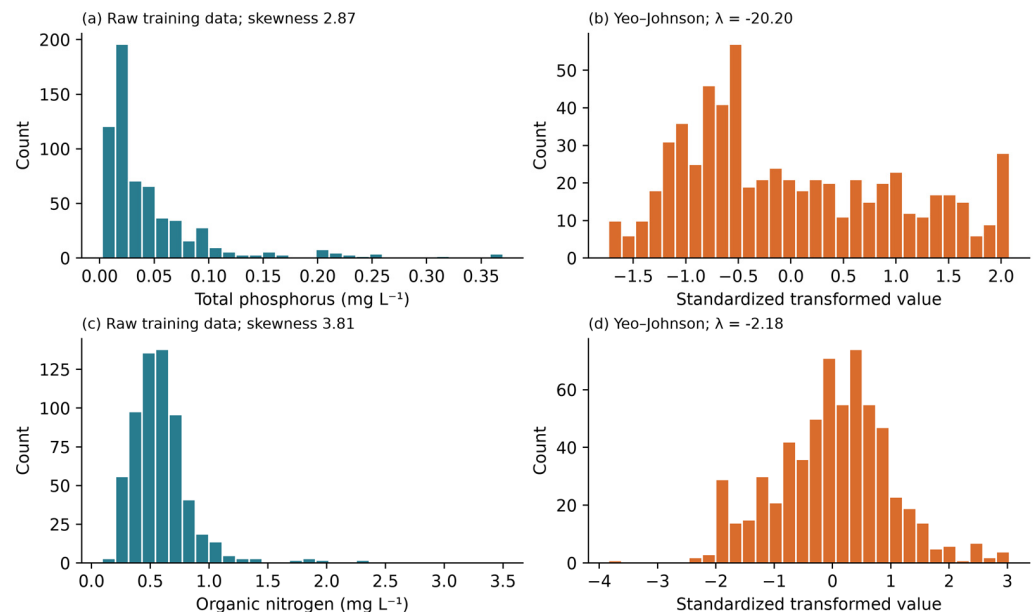


Figure 4. Distributions of (a,b) total phosphorus and (c,d) organic nitrogen before and after Yeo–Johnson transformation and standardization. Parameters were fitted to median-imputed 2014–2019 training data; the histograms display the nonmissing training observations. These diagnostic panels are not a separate validation test.

2.2. Machine Learning Methods

Random Forest averages predictions from decision trees trained on bootstrap samples, with random feature subsets considered during splitting [24]. XGBoost adds trees sequentially to reduce a regularized loss and uses shrinkage and subsampling to constrain model complexity [25,26]. These approaches have been applied to environmental tabular data, including water quality, lake level, lake temperature, and evapotranspiration [12,23,27,37,38]. Their relative performance depends on the task and dataset. LSTM networks use recurrent states to represent sequential dependencies [28–32]; however, the one-step implementation used here cannot learn relationships across successive sampling dates.

For the initial evaluations in Sections 3.1–3.5, Random Forest used 500 trees, square-root feature subsampling, and a minimum leaf size of three. XGBoost used 500 boosting rounds, a learning rate of 0.03, a maximum tree depth of three, a minimum child weight of three, row and column subsampling fractions of 0.85, an L2 regularization of 1.0, and squared-error loss. The LSTM received inputs of shape one time step by 43 predictors and used a 32-unit layer with tanh activation, a dropout rate of 0.20, and a dense output layer. It was trained with Adam (learning rate 0.01), mean squared error loss, batch size 32, and a maximum of 80 epochs. A 20% internal validation split controlled early stopping with a patience of eight, a minimum improvement of 10^{-4} , and restoration of the best weights. Preprocessing was fitted to the outer training set before this internal split. A base random seed of 42 was used; five-fold model runs used seeds 43–47. The evaluated configuration was a one-step benchmark because the sparse, irregular monitoring design did not provide the regular multistep inputs used in conventional recurrent forecasting. Python 3.12, NumPy 2.4.6, and TensorFlow 2.21.0 seeds were set for each model run using TensorFlow’s Keras seed utility. Deterministic TensorFlow operations were enabled, and the Keras session was cleared before each LSTM fit. The unit-corrected rerun used Python 3.12, scikit-learn 1.9.0, XGBoost 3.3.0, and TensorFlow 2.21.0; full software versions are recorded in Supplementary Code S1. Layer initializers were not overridden from the Keras defaults. Model settings were selected after exploratory comparisons of manually specified configurations on the available data, with RMSE and R^2 used in the exploratory

summaries. The available notebook does not document a prespecified search space, an exclusive selection metric, or a complete configuration-by-score history linking those trials to the final settings. No nested tuning and evaluation design or untouched final test set was reserved for the initial analysis. Those comparisons are therefore exploratory and may be optimistic because evaluation results informed configuration selection. They are not unbiased estimates of generalization performance; neither training-only preprocessing nor holding settings fixed during the rerun eliminates this selection bias. The documented settings were held fixed during the unit-corrected rerun; no additional hyperparameter search was performed in that rerun. The subsequent nested search and its distinct fixed-epoch LSTM training are described in Section 2.5.

2.3. Cross-Validation Tests for ML Method Evaluation

Three initial validation designs were applied consistently across models. Shuffled five-fold cross-validation divided the 1012 station–date records into five approximately equal folds using a random seed of 42. Each fold was held out once while the remaining four were used for training. Splitting was performed by observation, not by station; stations were not blocked in this shuffled design, and repeated observations from a station could occur in both training and test folds. Imputation, power transformation, and standardization were refitted on each training split. The chronological holdout used 631 observations from 2014–2019 for training and 381 from 2020–2024 for testing. Spatial validation withheld all observations from one focal lake and trained on all remaining records, repeating the procedure separately for the four focal lakes. The same model configurations were used in all three designs.

For the initial shuffled five-fold cross-validation, RMSE, MAE, and R^2 were calculated separately for each test fold; Table 4 reports means and standard deviations across folds. Metrics calculated from pooled out-of-fold predictions are reported in Section 3.3 and need not equal mean fold metrics. For uncertainty estimation, 5000 bootstrap samples were drawn by resampling the 33 monitoring stations with replacement (NumPy generator seed 42), retaining all station–date records from each sampled station. Identical resamples were used for all three models, and metrics and paired differences were recalculated on each resample. The 2.5th and 97.5th percentiles define the intervals reported in Section 3.2 and Table S2. These intervals are conditional on the fitted predictions; they do not include model refitting or tuning uncertainty, correct configuration-selection bias, or remove dependence between training and test records from shared stations. Resampling stations for uncertainty estimation is not station-blocked model validation. Station resampling retains within-station temporal dependence but does not account for dependence between stations in the same lake or common sampling periods. Intervals are pointwise and were not adjusted for multiple comparisons.

Table 4. Exploratory shuffled five-fold cross-validation metrics for within-network interpolation. Means and standard deviations (SDs) are calculated across folds. RMSE and MAE are in $\mu\text{g L}^{-1}$; R^2 is dimensionless. Stations were not blocked; shared-station dependence and configuration-selection bias remain.

Method	RMSE Mean	RMSE SD	MAE Mean	MAE SD	R^2 Mean	R^2 SD
Random Forest	11.54	3.21	4.36	1.08	0.385	0.207
XGBoost	10.45	3.50	4.12	0.98	0.484	0.254
LSTM	11.89	3.45	4.58	1.15	0.357	0.192

The chronological and spatial holdouts describe transfer under the specified split or excluded lake, whereas shuffled cross-validation describes interpolation within the sampled network. Shared stations and sampling periods across shuffled folds create dependence despite training-only preprocessing, so these metrics cannot establish performance at independent stations. Configuration-selection bias applies to all three evaluation designs. The chronological holdout uses contemporaneous predictors from 2020–2024 after training on 2014–2019; it assesses conditional concentration estimation for that split, not advance forecasting. This original holdout is interpreted only for that split; Section 2.5 describes the additional cutoffs. Neither assessment establishes overall future-prediction capability. Spatial validation comprised exactly four focal-lake holdouts, not an exhaustive leave-one-lake-out evaluation of the network.

2.4. Feature Importance Method

Feature importance was evaluated on the chronological holdout using the exact fitted chronological models and the same 2014–2019 training and 2020–2024 test periods. Each predictor was permuted in the test set while other columns were left unchanged, and RMSE was recalculated in the original Chl-a units. The increase relative to the unpermuted baseline was averaged over eight permutations; the standard deviation across permutations describes repeat-to-repeat variability. This model-agnostic measure permits comparison using the same error scale but measures reliance on a predictor rather than a causal effect. Correlated predictors can share information, and independent shuffling can produce combinations not observed in the original data. Rankings must also be interpreted in light of the weak baseline temporal performance. They do not establish the accuracy of a reduced monitoring or prediction system. A single NumPy random generator initialized with seed 42 generated successive permutations in the saved predictor-column order and model order Random Forest, XGBoost, and LSTM; it was not reset for each predictor.

2.5. Additional Station-Blocked Validation, Nested Configuration Selection, and Temporal Validation

A supplementary protocol of nested configuration selection was recorded before the additional model fits, after the initial analyses had been examined. Five GroupKFold outer splits partitioned the 33 stations without shuffling; all observations from a station were assigned to one assessment fold. Three GroupKFold inner splits were constructed using only the stations in each outer training set. The same inner and outer partitions were used for all models. Every inner fit estimated its own preprocessing parameters; the selected configuration was then refitted on the complete outer training set and evaluated once on the outer assessment fold. An accompanying fixed-configuration analysis used the same outer folds. All 1012 observations received one out-of-fold prediction in each design.

The candidate grid comprised Random Forest minimum leaf sizes {3, 1, 5}; XGBoost maximum depths {3, 2, 5}; and LSTM (units, learning rate) pairs {(32, 0.01), (16, 0.001), (64, 0.001)}. Other settings followed Section 2.2, except that all new LSTM fits used 80 fixed epochs with no early stopping or validation subset. The first listed candidate defined the fixed-configuration comparator. Selection minimized the arithmetic mean of the three inner-fold RMSE values in $\mu\text{g L}^{-1}$; ties favored the first candidate. This procedure constitutes nested configuration selection: hyperparameter choice uses only inner-fold data, while performance is assessed on outer-fold stations that played no role in selecting the configuration. This separation limits selection bias within the specified candidate grid, although historical exploratory choices and pooled predictor screening still limit independent interpretation. Outer assessment results were not used in this selection. This limited grid was fixed for the new analysis, not prespecified before historical exploration. It is not an exhaustive optimization of each architecture.

Within every inner, outer, or temporal training set, all-missing and constant predictors were removed before fitting median imputation, Yeo–Johnson transformation, and standardization. The response transformation and inverse transformation followed Section 2.1. The historical 43-variable availability screen was retained. Outer seeds were 42 plus the fold number; inner seeds were 1000 plus ten times the outer fold number plus the inner fold number. Seeds were shared across candidates and models. The additional runtime used the same core library versions listed in Section 2.2; exact dependency versions are archived in Supplementary Code S2.

Temporal sensitivity was assessed with six expanding training periods beginning in 2014 and ending in 2017, 2018, 2019, 2020, 2021, or 2022. Each model was assessed on the next two calendar years, using the fixed configurations above, training-only preprocessing, and a seed of 42 plus the cutoff year. No future records were used for fitting or stopping. Adjacent assessment windows overlap by one year and are not independent replicates. These tests use contemporaneous water quality inputs; they assess conditional estimation in later periods rather than forecasts issued in advance.

RMSE was the primary selection and reporting metric; MAE and R^2 were secondary. A training-set arithmetic-mean predictor was assessed on each outer and temporal test set as a deployable reference. For conditional uncertainty, 5000 shared station-bootstrap resamples (seed 42) were applied to the fixed pooled predictions, with percentile intervals and paired RMSE differences. These pointwise intervals do not refit models or include configuration-selection uncertainty. Complete fold scores, candidate scores, predictions, and temporal results are retained in Supplementary Code S2.

3. Results

3.1. Lake-Specific Time-Series Predictions

Figure 5 compares observed Chl-a with predictions from Random Forest, XGBoost, and LSTM over time in the four focal lakes. The plots provide an initial view of where the methods give similar estimates and where they differ from observations. All three methods underestimate the largest same-date lake-mean concentrations in Munson and Iamonia, whereas Jackson and Bradford have lower observed concentration ranges. These are the shuffled five-fold out-of-fold predictions: each station–date observation was excluded from the fit used to predict it, although other observations from the same station or lake could remain in training. The plots consequently describe exploratory interpolation with within-station dependence across folds. Values are averaged across stations within each lake on the same date for display; quantitative metrics use station–date records before averaging.

Visual agreement at lower concentrations alone cannot establish which method performs better or whether performance transfers to later years or an excluded lake. The following sections therefore quantify model differences using five-fold error metrics and paired station-bootstrap intervals, then examine chronological and focal-lake holdouts. Section 3.6 then reports station-blocked, nested, and repeated temporal assessments. The initial evaluations describe dependent within-network interpolation, transfer for the single 2014–2019 versus 2020–2024 split, and transfer to the four withheld lakes; all remain exploratory because of configuration selection. The plots alone do not establish that adding high-concentration training observations would resolve the observed errors.

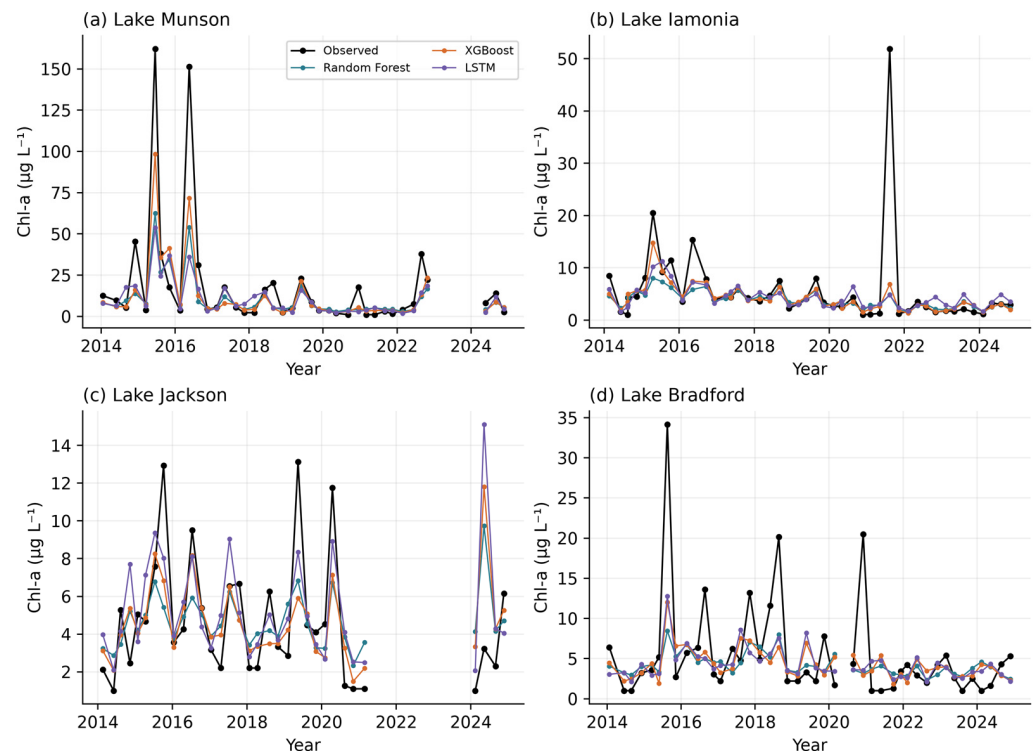


Figure 5. Observed and out-of-fold predicted Chl-a concentrations over time for (a) Lake Munson, (b) Lake Iamonia, (c) Lake Jackson, and (d) Lake Bradford. Predictions from Random Forest, XGBoost, and LSTM come from shuffled five-fold cross-validation. Points show same-date lake means of station-level observations or predictions. Lines are visual guides and are broken when successive available observations are more than 180 days apart; no observations are interpolated. These exploratory interpolation plots are not full-data fits or advance forecasts. Stations were not blocked; shared stations across folds and configuration selection limit interpretation as generalization performance.

3.2. Shuffled Five-Fold Cross-Validation

In this exploratory comparison, XGBoost had the lowest mean five-fold RMSE ($10.45 \mu\text{g L}^{-1}$) and MAE ($4.12 \mu\text{g L}^{-1}$), and the highest mean R^2 (0.484), followed by Random Forest and LSTM (Table 4). Fold-to-fold variability was substantial. Stations were not blocked, so repeated observations from the same station remained across training and test folds. These metrics characterize within-network interpolation and may be optimistic because of this dependence and configuration-selection bias. Figure 6a shows mean RMSE with one standard deviation across folds; Figure 6d shows pooled out-of-fold R^2 with station-bootstrap intervals. The paired pooled RMSE difference for XGBoost minus Random Forest was $-0.98 \mu\text{g L}^{-1}$ (95% interval -1.93 to -0.15), and that for XGBoost minus LSTM was $-1.38 \mu\text{g L}^{-1}$ (-2.85 to -0.15 ; Table S2). These conditional intervals support a numerical advantage within the shuffled predictions but do not correct either source of optimism or establish superiority at independent stations, in other time periods, or across lakes.

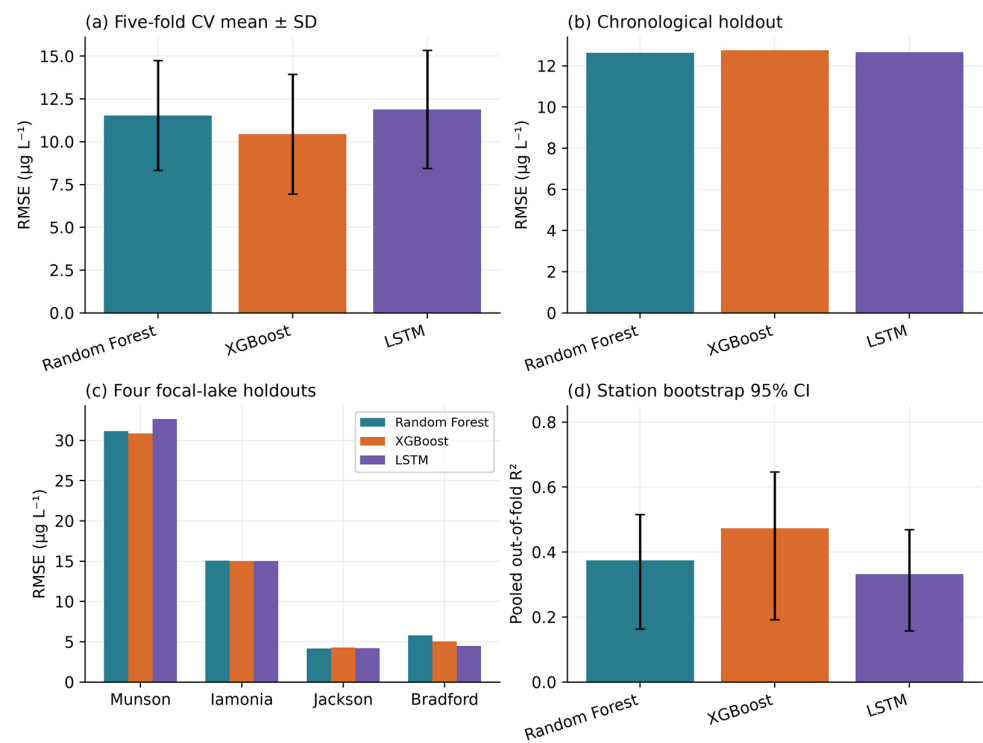


Figure 6. Exploratory model performance under (a) shuffled five-fold validation (mean RMSE $\pm$ fold SD), (b) the single 2014–2019 training versus 2020–2024 chronological holdout, and (c) four focal-lake holdouts. Panel (d) shows pooled out-of-fold R^2 with 95% station-bootstrap percentile intervals from 5000 resamples; paired differences are reported in Table S2. RMSE is in $\mu\text{g L}^{-1}$. Stations were not blocked in panels (a,d). Bootstrap intervals condition on existing predictions and do not remove training–test dependence or configuration-selection bias. Panel (b) is specific to the stated split; additional station-blocked and temporal assessments are reported separately in Section 3.6.

3.3. Chronological and Spatial Cross-Validation

For the original 2014–2019 training versus 2020–2024 test split, chronological-holdout performance was similar across models (Table 5; Figure 6b). RMSE ranged from 12.64 to 12.76 $\mu\text{g L}^{-1}$, MAE from 4.24 to 4.36 $\mu\text{g L}^{-1}$, and R^2 from 0.054 to 0.071. These R^2 values indicate little improvement over the test-set mean as a descriptive squared-error reference. The numerically better shuffled results, which retain within-station dependence, were not reproduced in this particular 2020–2024 holdout. This contrast is specific to the stated holdout; the additional cutoffs are examined separately in Section 3.6. These results also remain subject to configuration-selection bias. Changes in sampled conditions and the limited representation of extreme observations are possible explanations, but their individual contributions were not isolated by this analysis.

Table 5. Chronological-holdout performance after training on 2014–2019 and testing on 2020–2024. Counts are station–date observations. RMSE and MAE are in $\mu\text{g L}^{-1}$. This is the original chronological split; additional cutoffs appear in Section 3.6. Historical configuration-selection bias remains.

Method	Training Records	Test Records	RMSE	MAE	R^2
Random Forest	631	381	12.64	4.24	0.071
XGBoost	631	381	12.76	4.36	0.054
LSTM	631	381	12.67	4.32	0.067

Figure 7 compares observed and predicted concentrations for the pooled five-fold predictions (panels a–c) and the single 2014–2019 training versus 2020–2024 holdout

(panels d–f). The pooled five-fold metrics differ from the means in Table 4 because the errors and total variation are aggregated differently. The shuffled panels retain within-station training–test dependence, and all panels remain subject to configuration-selection bias. All models showed substantial underestimation of the largest observed concentrations, especially in the chronological holdout. More accurate estimates at low-to-moderate concentrations should therefore not be interpreted as reliable performance across the full concentration range.

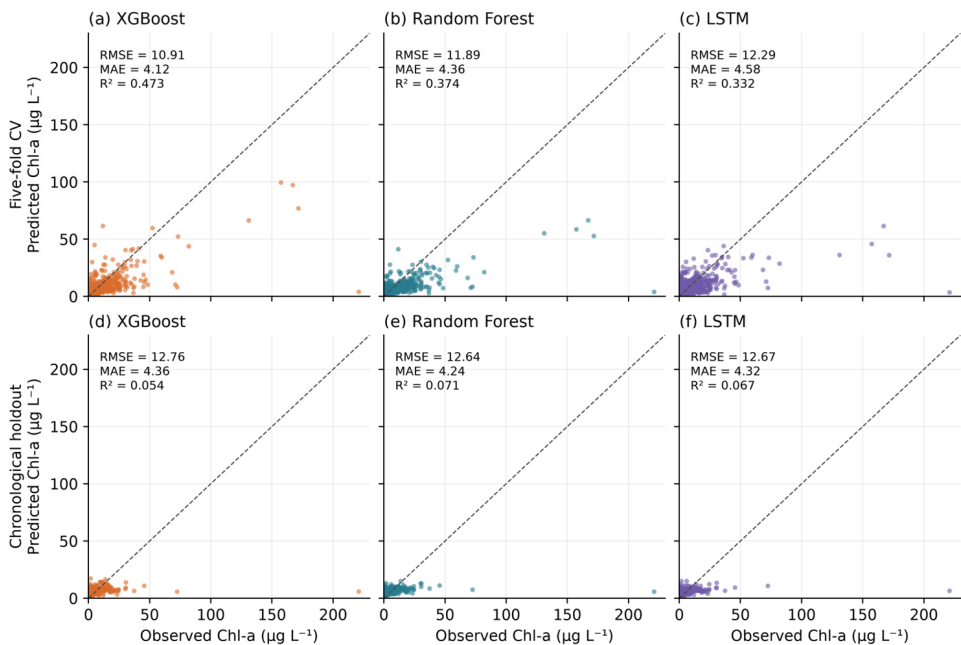


Figure 7. Observed and predicted Chl-a for (a–c) pooled shuffled five-fold out-of-fold predictions and (d–f) the single 2014–2019 training versus 2020–2024 holdout. Model order is XGBoost, Random Forest, and LSTM. Dashed lines indicate equality. RMSE and MAE are in $\mu\text{g L}^{-1}$, and displayed metrics are calculated from the points in each panel. Stations were not blocked in (a–c). All metrics are exploratory and subject to configuration-selection bias; (d–f) do not establish performance across other time periods.

Spatial performance varied among the four withheld lakes (Table 6; Figure 6c). RMSE was highest for Lake Munson ($30.84\text{--}32.65\ \mu\text{g L}^{-1}$), compared with $4.17\text{--}4.28\ \mu\text{g L}^{-1}$ for Lake Jackson. Lake Iamonia had RMSE near $15\ \mu\text{g L}^{-1}$ and R^2 of only $0.022\text{--}0.027$. Absolute errors were lower for Lakes Jackson and Bradford, but this partly reflects their lower concentration ranges; R^2 differed among models, particularly at Bradford ($0.160\text{--}0.500$). Thus, lower RMSE alone does not establish satisfactory transfer. The results show limited transfer to high-concentration conditions represented in Munson and Iamonia, without implying that such conditions are universally impossible to model.

Table 6. Leave-one-lake-out validation for the four focal lakes. All records from the named lake were excluded from training. RMSE and MAE are in $\mu\text{g L}^{-1}$; counts are station–date observations.

Held-Out Lake	Method	Training Records	Test Records	RMSE	MAE	R^2
Lake Munson	Random Forest	934	78	31.15	13.10	0.192
Lake Munson	XGBoost	934	78	30.84	13.23	0.207
Lake Munson	LSTM	934	78	32.65	13.23	0.112
Lake Iamonia	Random Forest	760	252	15.06	3.53	0.022
Lake Iamonia	XGBoost	760	252	15.02	3.39	0.027

Table 6. Cont.

Held-Out Lake	Method	Training Records	Test Records	RMSE	MAE	R ²
Lake Iamonia	LSTM	760	252	15.03	3.48	0.026
Lake Jackson	Random Forest	889	123	4.17	2.87	0.239
Lake Jackson	XGBoost	889	123	4.28	2.83	0.198
Lake Jackson	LSTM	889	123	4.20	2.64	0.229
Lake Bradford	Random Forest	969	43	5.79	3.27	0.160
Lake Bradford	XGBoost	969	43	5.02	3.19	0.370
Lake Bradford	LSTM	969	43	4.47	3.03	0.500

3.4. Lake-Specific Prediction Errors

The pooled out-of-fold predictions were further examined by lake to assess where interpolation errors occurred (Figure 8). Each displayed prediction was generated while that observation was excluded from model fitting, although other observations from the same lake or station could be present in training. These plotted predictions were not station-blocked, so within-station dependence remains; the plots are also subject to configuration-selection bias. These are the same five-fold predictions summarized in Table 4 and Figure 7, not an additional validation experiment. All methods underestimated the largest observation in Lake Iamonia. Lake Munson showed a wider spread of errors than Jackson and Bradford, consistent with its broader concentration range.

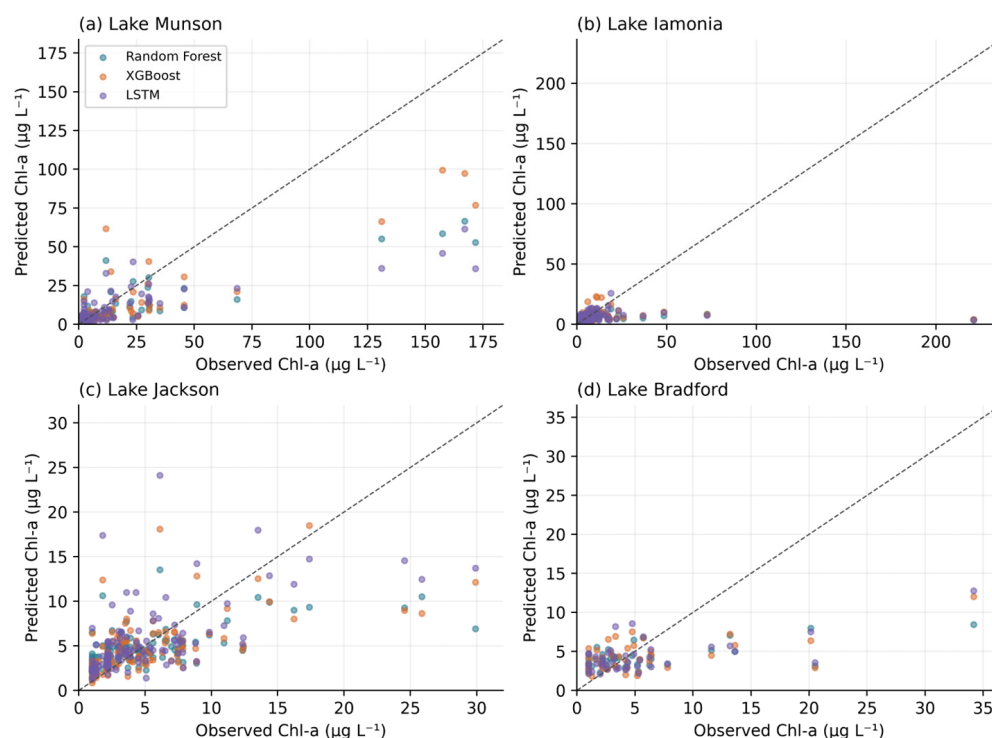


Figure 8. Lake-specific observed and predicted Chl-a from shuffled five-fold out-of-fold predictions: (a) Munson, (b) Iamonia, (c) Jackson, and (d) Bradford. Each prediction excludes its own observation but may share stations across training and test folds. These exploratory interpolation results are not station-blocked, leave-one-lake-out, or chronological predictions and remain subject to configuration-selection bias. Dashed lines indicate equality.

3.5. Feature Importance

Total phosphorus had the largest mean temporal permutation importance for Random Forest, and biochemical oxygen demand ranked first for XGBoost. Total Kjeldahl

nitrogen (TKN) and total suspended solids had nearly equal leading importance values for LSTM (Figure 9). The largest mean RMSE increases were approximately 0.060, 0.065, and 0.067 $\mu\text{g L}^{-1}$ for Random Forest, XGBoost, and LSTM, respectively. These increases were small compared with baseline RMSE values of 12.64–12.76 $\mu\text{g L}^{-1}$, reinforcing the need for cautious interpretation. Figure 10 provides a separate descriptive analysis of Chl-a and six selected measured variables. Total phosphorus had the largest positive linear correlation with Chl-a among those shown ($r = 0.43$). This association does not validate the model-importance rankings or establish causality.

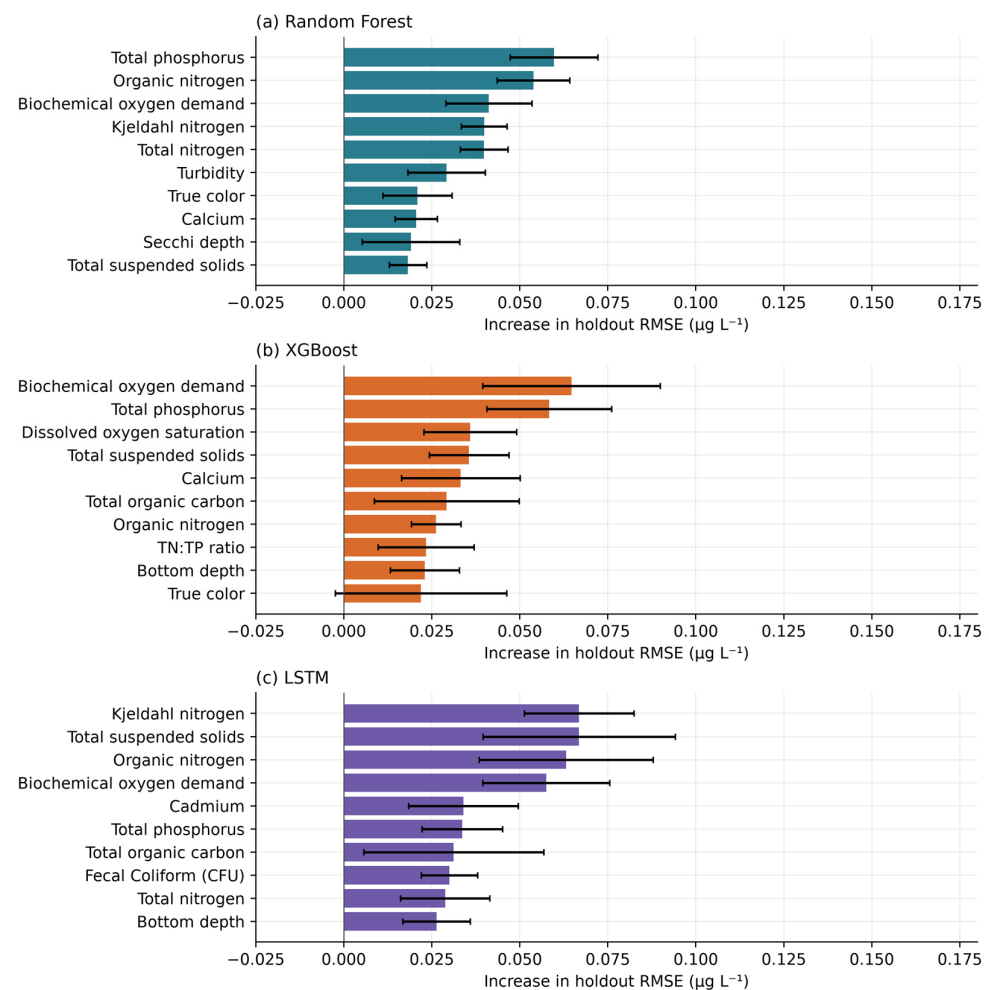


Figure 9. Ten leading predictors for each model based on permutation importance in the 2020–2024 holdout. Bars show the mean increase in RMSE across eight permutations; error bars show ± 1 SD across permutations, not confidence intervals. All panels use the same horizontal scale. The figure label “Kjeldahl nitrogen” denotes total Kjeldahl nitrogen (TKN).

The rankings differed across models. Random Forest emphasized total phosphorus, organic nitrogen, biochemical oxygen demand, total Kjeldahl nitrogen (TKN), and total nitrogen. XGBoost emphasized biochemical oxygen demand, total phosphorus, oxygen saturation, suspended solids, and calcium. LSTM emphasized total Kjeldahl nitrogen (TKN), suspended solids, organic nitrogen, biochemical oxygen demand, and cadmium. Total phosphorus and biochemical oxygen demand appeared among the ten leading predictors for all three models. Because LSTM received a single station–date vector, its rankings describe sensitivity to contemporaneous inputs and cannot be attributed to learned long-term temporal dynamics.

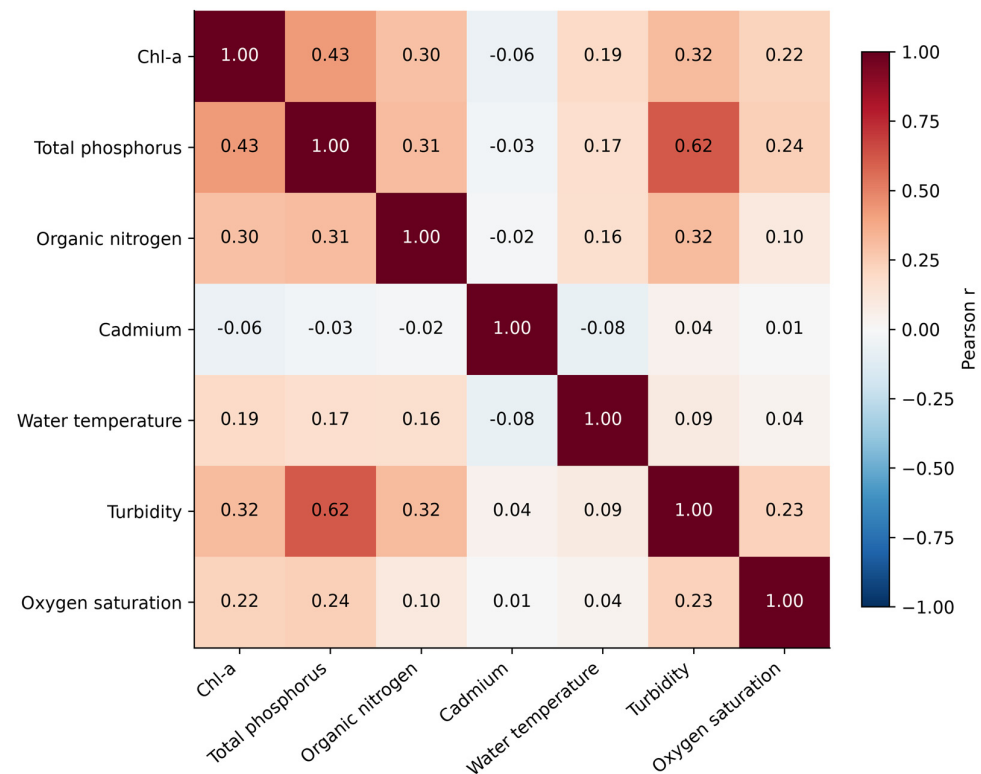


Figure 10. Pearson correlations among Chl-a and six selected measured variables: total phosphorus, organic nitrogen, cadmium, water temperature, turbidity, and dissolved oxygen saturation. Correlations use raw values and pairwise complete observations ($N = 1007$ – 1012). The variables provide descriptive context and are not the six highest-ranked predictors in Figure 9.

3.6. Station-Blocked Validation, Nested Configuration Selection, and Repeated Temporal Validation

Table 7 pools the station-held-out predictions before calculating metrics. The five outer station folds had no shared stations between training and assessment sets. Fixed-configuration pooled RMSE values were 12.08 , 10.70 , and $10.48 \mu\text{g L}^{-1}$ for Random Forest, XGBoost, and LSTM, respectively. Nested selection yielded 11.92 , 10.49 , and $10.91 \mu\text{g L}^{-1}$, with pooled R^2 values of 0.371 , 0.513 , and 0.473 (Table 7). These pooled metrics weight station–date observations equally; they are not averages of station or fold metrics. The training-mean reference had an RMSE of $15.13 \mu\text{g L}^{-1}$. All candidate-fold scores and selected configurations are provided in Supplementary Code S2.

For nested station-held-out predictions, the paired RMSE difference for XGBoost minus LSTM was $-0.42 \mu\text{g L}^{-1}$ (95% station-bootstrap interval -2.18 to 0.83). These intervals are conditional on the existing predictions; they do not establish a general model ranking. Within-fold station exclusion addresses the original overlap concern, but shared lakes and sampling periods remain possible across folds.

Across the six two-year temporal assessment windows, R^2 ranged from 0.024 to 0.410 for Random Forest, -0.010 to 0.340 for XGBoost, and -5.286 to 0.016 for LSTM (Table 8). Random Forest had the lowest RMSE among the three models in all six windows; LSTM had negative R^2 in four. Both the observed concentration distribution and the amount of training data change with the cutoff. The original five-year 2020–2024 holdout in Table 5 is a separate assessment; its values should not be substituted for the two-year 2020–2021 window. All six planned cutoffs are reported, including less favorable results. These retrospective tests characterize the evaluated periods without establishing performance in other future conditions.

Table 7. Pooled station-held-out performance from five outer folds, with and without nested configuration selection. Each design includes all 1012 observations from 33 stations. RMSE and MAE are in $\mu\text{g L}^{-1}$; brackets are conditional 95% station-bootstrap intervals for RMSE. Fixed and nested fits use the same station folds and fixed 80-epoch LSTM training. Historical exploration and pooled predictor screening remain limitations.

Selection	Method	RMSE [95% Interval]	MAE	R ²
Fixed	Random Forest	12.08 [7.40, 16.24]	4.50	0.354
Fixed	XGBoost	10.70 [6.59, 14.91]	4.10	0.493
Fixed	LSTM	10.48 [6.68, 14.83]	4.26	0.514
Fixed	Training mean	15.13 [9.71, 19.72]	7.48	−0.013
Nested	Random Forest	11.92 [7.30, 16.10]	4.45	0.371
Nested	XGBoost	10.49 [6.49, 14.74]	4.06	0.513
Nested	LSTM	10.91 [7.07, 15.20]	4.39	0.473
Nested	Training mean	15.13 [9.71, 19.72]	7.48	−0.013

Table 8. Repeated chronological assessment using expanding training periods and the following two calendar years for testing. Entries are RMSE in $\mu\text{g L}^{-1}$, with R² in parentheses. *N* gives training–test station–date counts. Fixed configurations were used without selection on these assessment windows. Adjacent test windows overlap; the six results are not independent replicates. The reference predicts the training-set mean.

Training	Test	<i>N</i>	Random Forest	XGBoost	LSTM	Mean Reference
2014–2017	2018–2019	446/185	5.68 (0.410)	6.01 (0.340)	18.53 (−5.286)	8.28 (−0.256)
2014–2018	2019–2020	538/175	6.61 (0.254)	6.88 (0.192)	9.12 (−0.424)	8.13 (−0.131)
2014–2019	2020–2021	631/157	19.04 (0.024)	19.19 (0.009)	19.13 (0.016)	19.51 (−0.024)
2014–2020	2021–2022	713/153	19.04 (0.034)	19.24 (0.014)	19.24 (0.013)	19.56 (−0.019)
2014–2021	2022–2023	788/144	4.84 (0.370)	5.32 (0.240)	9.98 (−1.681)	7.14 (−0.371)
2014–2022	2023–2024	866/146	3.80 (0.174)	4.20 (−0.010)	7.15 (−1.925)	5.88 (−0.981)

4. Discussion

XGBoost had the best observed average performance in the exploratory shuffled five-fold comparison. This numerical ranking describes interpolation within the sampled network under the evaluated configurations; shared-station dependence and configuration-selection bias prevent interpreting it as an unbiased ranking of generalization performance. The additional nested station-held-out assessment produced pooled RMSE values of 11.92, 10.49, and 10.91 $\mu\text{g L}^{-1}$ for Random Forest, XGBoost, and LSTM, respectively (Table 7). Model rankings should consequently be interpreted within each validation design. Prior environmental applications provide context for their use [12,23–27], but the present results do not demonstrate that tree methods generally outperform sequence-based deep learning. The LSTM benchmark used only one time step, whereas applications of recurrent networks usually exploit sequential inputs [20,28–32,39]. The comparison therefore concerns the specific configurations evaluated here. It does not isolate the effects of architecture, hyperparameter selection, or sequence length.

The dataset comparisons in Table 1 highlight both record count and sampling structure. Large regional and national compilations can cover a wider range of conditions, but individual sites may remain sparsely sampled. Zhi et al. [20], for example, paired sparse chemistry data with daily hydrometeorological inputs. The present models used contemporaneous water quality measurements and lacked comparable daily driver sequences. Differences in inputs, targets, and validation designs preclude attributing differences between studies solely to sample size.

Temporal and spatial transfer impose the strongest constraints on interpretation. Chronological R^2 values were near zero for the original 2014–2019 training versus 2020–2024 test split. The moderate shuffled-validation metrics reflect dependent within-network interpolation. The six additional chronological assessments show how performance changes with the cutoff and assessment window (Table 8). Their overlapping test periods and changing concentration distributions preclude treating them as independent replicates or attributing RMSE changes solely to training-set size. They provide evidence for the tested periods, not a guarantee of performance in unobserved future conditions. Withholding a focal lake also produced substantial errors, especially at Munson and Iamonia. The lower absolute errors at Jackson and Bradford do not establish general transferability because their concentration distributions differ. Model choice alone did not resolve these limitations. Before operational application, validation should reflect the intended use, including the availability of predictor measurements and the range of conditions expected at the target lake.

The models estimate Chl-a conditional on water quality measurements collected at the same station and date. They do not provide an economical replacement for direct Chl-a measurement merely because they can combine 43 existing predictors. Their present use is a comparison of empirical associations and estimation where accompanying measurements already exist; a reduced-input model would require a separate validation. Calendar date, season, lags, and antecedent meteorological variables were not included. Concurrent water temperature may contain seasonal information, but it does not replace explicit seasonal or temporal validation. The plotted time axis is therefore a display coordinate, not an input sequence learned by the models.

Phosphorus and organic nitrogen can covary with algal biomass, while biochemical oxygen demand and organic carbon reflect organic matter. Suspended solids and turbidity can include algal cells as well as mineral particles and may therefore respond to the same biomass represented by Chl-a. Their importance cannot be interpreted solely as an effect of sediment transport or light limitation. The pooled models and permutation analysis do not separate within-lake from between-lake relationships; the LSTM ranking of suspended solids does not establish a distinct within-lake driver. A hierarchical or explicitly decomposed analysis would be needed for that inference.

Higher Chl-a concentrations accompanied by greater variability can increase uncertainty near trophic-class boundaries. Systematic underprediction of high values could assign an overly favorable class to an affected sample. Repeatedly high measured concentrations provide stronger evidence of high algal biomass than isolated model estimates, but classification also depends on the chosen trophic framework, sampling period, and other indicators [5,6]. This study did not evaluate classification accuracy or assign regulatory water-quality status, so higher variability should not be treated as proof of a particular quality class.

Lake-specific shuffled out-of-fold plots (Figures 5 and 8) remain subject to station and temporal dependence across training and test folds. The new grouped validation removes shared stations within each split, but stations from the same lake and observations from common periods can still occur on both sides; it is not lake-blocked or jointly space–time-blocked validation. Station-bootstrap intervals are conditional on fitted predictions and do not capture dependence between stations or refitting and selection uncertainty. The initial configurations were chosen after exploratory manual comparisons, without nested selection or an untouched final test set. The new inner-fold search avoids using its outer assessment outcomes for configuration selection, but cannot remove historical knowledge of the dataset or the pooled availability screen that defined the 43-predictor panel. The limited candidate grid, fixed random seeds, and single outer station partition also restrict

inference. New LSTM fits used fixed 80-epoch training rather than the original early stopping, so differences between old and new LSTM results cannot be attributed solely to validation design. The six additional temporal windows overlap, and their fixed settings retain historical selection bias. Only four focal-lake holdouts were evaluated. This study therefore remains exploratory; a fully specified pipeline should be tested on independently collected data before deployment.

5. Conclusions

This exploratory study evaluated Random Forest, XGBoost, and a one-step LSTM network using 1012 station–date observations and 43 water quality predictors from Leon County during 2014–2024. The four focal lakes contributed 43–252 observations each. Source units were harmonized before aggregation. Training-set median imputation, Yeo–Johnson transformation, and standardization were applied to predictors, and a standardized $\ln(1 + \text{Chl-a})$ response was used for fitting. This study compared shuffled interpolation, transfer for the original 2014–2019 versus 2020–2024 split, and transfer to a withheld focal lake with shared observation splits across methods. Additional tests excluded entire stations, nested configuration selection within training data, and repeated chronological assessment across six cutoffs.

XGBoost had the lowest observed mean shuffled five-fold RMSE ($10.45 \mu\text{g L}^{-1}$), but those predictions retain within-station training–test dependence. In the new nested station-blocked assessment, pooled RMSE was 11.92, 10.49, and $10.91 \mu\text{g L}^{-1}$ for Random Forest, XGBoost, and LSTM, respectively. These estimates assess the specified selection procedure on held-out stations within the existing network; they do not validate transfer to independent lakes. Random Forest had the lowest RMSE across all six additional temporal windows, whereas LSTM had negative R^2 in four. These findings complement the original 2014–2019 versus 2020–2024 result and show that the station-held-out ranking does not persist across all assessment designs. The analyses do not establish reliable advance forecasting. Historical manual selection and pooled predictor screening still limit independent assessment despite the new nested search. Conclusions remain specific to the evaluated data, configurations, and sampling designs, and the one-step LSTM results should not be generalized to multistep recurrent architectures.

Temporal-holdout permutation importance identified phosphorus and biochemical oxygen demand among the leading predictors across models, although rankings differed and the absolute error increases were small. These associations may inform further investigation but do not justify causal claims or removal of monitoring variables without additional validation. Future work should improve coverage of high-concentration events, obtain independent data for externally validated model selection, and evaluate rainfall, antecedent runoff, lake stage, land cover, station location, and seasonal predictors.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/limnolrev26040064/s1>. The accompanying Supplementary_Tables.docx contains Table S1, listing the 43 predictors and their missingness, and Table S2, reporting paired differences in pooled out-of-fold performance with station-bootstrap intervals. Supplementary_Code_S1.zip provides the reproducible unit-corrected analysis and its outputs. Supplementary_Code_S2.zip contains the additional validation protocol recorded before the new fits, corrected input matrices, model-fitting and verification scripts, candidate scores, split membership, predictions, uncertainty summaries, and software versions.

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 published 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 records were obtained from the Leon County lake monitoring program. Supplementary Code S1 contains the unit-harmonization and evaluation scripts, the source-data snapshot, the conversion audit, software versions, and the numerical outputs used in this revision. The public project repository is <https://github.com/alexwang1500/LeonCountyCode> (accessed on 15 October 2026); Supplementary Code S1 records the specific analysis used for the revised results. Supplementary Code S2 records the additional station-blocked, nested, and repeated temporal analyses reported in Section 3.6.

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 conflicts 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. [CrossRef]
2. Paerl, H.W.; Huisman, J. Climate change: A catalyst for global expansion of harmful cyanobacterial blooms. *Environ. Microbiol. Rep.* **2009**, *1*, 27–37. [CrossRef]
3. Heisler, J.; Glibert, P.; Burkholder, J.; Anderson, D.; Cochlan, W.; Dennison, W.; Dortch, Q.; Gobler, C.; Heil, C.; Humphries, E.; et al. Eutrophication and harmful algal blooms: A scientific consensus. *Harmful Algae* **2008**, *8*, 3–13. [CrossRef]
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. [CrossRef]
6. OECD. *Eutrophication of Waters: Monitoring, Assessment and Control*; Organisation for Economic Co-Operation and Development: Paris, France, 1982.
7. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. [CrossRef]
8. Shen, C. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resour. Res.* **2018**, *54*, 8558–8593. [CrossRef]
9. 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. [CrossRef]
10. Wu, N.; Huang, J.; Schmalz, B.; Fohrer, N. Modeling daily chlorophyll-a dynamics in a German lowland river using artificial neural networks and multiple linear regression approaches. *Limnology* **2014**, *15*, 47–56. [CrossRef]
11. Park, Y.; Cho, K.H.; Park, J.; Cha, S.M.; Kim, J.H. 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. [CrossRef]
12. Chen, K.; Chen, H.; Zhou, C.; Huang, Y.; Qi, X.; Shen, R.; Liu, F.; Zuo, M.; Zou, X.; Wang, J.; et al. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Res.* **2020**, *171*, 115454. [CrossRef]
13. 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. [CrossRef]
14. Pahlevan, N.; Smith, B.; Schalles, J.; Binding, C.; Cao, Z.; Ma, R.; Alikas, K.; Kangro, K.; Gurlin, D.; Hà, N.; et al. Seamless retrievals of chlorophyll-a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach. *Remote Sens. Environ.* **2020**, *240*, 111604. [CrossRef]
15. 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. [CrossRef]
16. Gholizadeh, M.H.; Melesse, A.M.; Reddi, L. A comprehensive review on water quality parameters estimation using remote sensing techniques. *Sensors* **2016**, *16*, 1298. [CrossRef]
17. Glasgow, H.B.; Burkholder, J.M.; Reed, R.E.; Lewitus, A.J.; Kleinman, J.E. 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. [CrossRef]
18. Kim, Y.H.; Im, J.; Ha, H.K.; Choi, J.-K.; Ha, S. Machine learning approaches to coastal water quality monitoring using GOCI satellite data. *GISci. Remote Sens.* **2014**, *51*, 158–174. [CrossRef]

19. Hafeez, S.; Wong, M.S.; Ho, H.C.; Nazeer, M.; Nichol, J.E.; Abbas, S.; Tang, D.; Lee, K.-H.; Pun, L. 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. [CrossRef]
20. Zhi, W.; Feng, D.; Tsai, W.-P.; Sterle, G.; Harpold, A.; Shen, C.; Li, L. 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. [CrossRef]
21. 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. [CrossRef]
22. 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. [CrossRef]
23. 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. [CrossRef]
24. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
25. 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.
26. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **2021**, *54*, 1937–1967. [CrossRef]
27. Huang, G.; Wu, L.; Ma, X.; Zhang, W.; Fan, J.; Yu, X.; Zeng, W.; Zhou, H. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. *J. Hydrol.* **2019**, *574*, 1029–1041. [CrossRef]
28. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
29. Kratzert, F.; Klotz, D.; Brenner, C.; Schulz, K.; Herrnegger, M. Rainfall-runoff modelling using Long Short-Term Memory networks. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 6005–6022. [CrossRef]
30. Hu, C.; Wu, Q.; Li, H.; Jian, S.; Li, N.; Lou, Z. Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. *Water* **2018**, *10*, 1543. [CrossRef]
31. Zhou, J.; Wang, Y.; Xiao, F.; Wang, Y.; Sun, L. Water quality prediction method based on IGRA and LSTM. *Water* **2018**, *10*, 1148. [CrossRef]
32. 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. [CrossRef]
33. Leon County. Waterbody Summary 2019 Lake Iamonia. 2019. Available online: <https://cms.leoncountyfl.gov/Portals/0/publicworks/engservices/docs/WQdata/Current/Iamonia%20Basin/Waterbody%20Summaries%202019%20Lake%20Iamonia.pdf> (accessed on 15 October 2026).
34. Florida Department of Environmental Protection. Lake Jackson Aquatic Preserve Management Plan. 2020. Available online: <https://floridadep.gov/rcp/aquatic-preserve/documents/lake-jackson-aquatic-preserve-management-plan> (accessed on 23 September 2026).
35. Leon County Water Resources. Lake Munson. Available online: <https://www.leoncountywater.org/Protecting-and-Preserving-Our-Water-Quality/Lake-Munson> (accessed on 23 September 2026).
36. Yeo, I.K.; Johnson, R.A. A new family of power transformations to improve normality or symmetry. *Biometrika* **2000**, *87*, 954–959. [CrossRef]
37. Zhu, S.; Hrnjica, B.; Ptak, M.; Choiński, A.; Sivakumar, B. Forecasting of water level in multiple temperate lakes using machine learning methods. *J. Hydrol.* **2020**, *585*, 124819. [CrossRef]
38. Tiyasha, T.; Tung, T.M.; Yaseen, Z.M. A survey on river water quality modelling using artificial intelligence methods: 2000–2020. *J. Hydrol.* **2020**, *585*, 124670. [CrossRef]
39. 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. [CrossRef]

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