AGU ADVANCING EARTH AND SPACE SCIENCES

Earth's Future

RESEARCH ARTICLE
10.1029/2024EF004493

Estimates of Lake Nitrogen, Phosphorus, and Chlorophyll-*a* Concentrations to Characterize Harmful Algal Bloom Risk Across the United States

Meredith M. Brehob[1], Michael J. Pennino[2], Amalia M. Handler[3], Jana E. Compton[3], Sylvia S. Lee[4], and Robert D. Sabo[2]

[1]Oak Ridge Institute for Science and Education (ORISE), U.S. EPA, Office of Research and Development, Center for Public Health and Environmental Assessment, Health & Environmental Effects Assessment Division, Washington, DC, USA, [2]U.S. EPA, Office of Research and Development, Center for Public Health and Environmental Assessment, Health & Environmental Effects Assessment Division, Washington, DC, USA, [3]U.S. EPA, Office of Research and Development, Center for Public Health and Environmental Assessment, Pacific Ecological Systems Division, Corvallis, OR, USA, [4]U.S. EPA, Office of Research and Development, Center for Public Health and Environmental Assessment, Integrated Climate Sciences Division, Washington, DC, USA

Key Points:

- Chlorophyll-*a* and nutrient concentrations in ~112,000 lakes were predicted using widely available national data sets and machine learning
- Nutrients far outweigh other environmental predictors in driving chlorophyll-*a* concentrations in lakes
- With more information about likely chlorophyll-*a* concentrations, managers can prioritize lakes at risk for harmful algal bloom production

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

M. J. Pennino,
pennino.michael@epa.gov

Citation:

Brehob, M. M., Pennino, M. J., Handler, A. M., Compton, J. E., Lee, S. S., & Sabo, R. D. (2024). Estimates of lake nitrogen, phosphorus, and chlorophyll-*a* concentrations to characterize harmful algal bloom risk across the United States. *Earth's Future*, *12*, e2024EF004493. https://doi.org/10.1029/2024EF004493

Received 30 JAN 2024
Accepted 17 JUN 2024

Abstract Excess nutrient pollution contributes to the formation of harmful algal blooms (HABs) that compromise fisheries and recreation and that can directly endanger human and animal health via cyanotoxins. Efforts to quantify the occurrence, drivers, and severity of HABs across large areas is difficult due to the resource intensive nature of field monitoring of lake nutrient and chlorophyll-*a* concentrations. To better characterize how nutrients interact with other environmental factors to produce algal blooms in freshwater systems, we used spatially explicit and temporally matched climate, landscape, in-lake characteristic, and nutrient inventory data sets to predict nutrients and chlorophyll-*a* across the conterminous US (CONUS). Using a nested modeling approach, three random forest (RF) models were trained to explain the spatiotemporal variation in total nitrogen (TN), total phosphorus (TP), and chlorophyll-*a* concentrations across US EPA's National Lakes Assessment ($n = 2{,}062$). Concentrations of TN and TP were the most important predictors and, with other variables, the RF model accounted for 68% of variation in chlorophyll-*a*. We then used these RF models to extrapolate lake TN and TP predictions to lakes without nutrient observations and predict chlorophyll-*a* for ~112,000 lakes across the CONUS. Risk for high chlorophyll-*a* concentrations is highest in the agriculturally dominated Midwest, but other areas of risk emerge in nutrient pollution hot spots across the country. These catchment and lake-specific results can help managers identify potential nutrient pollution and chlorophyll-*a* hot spots that may fuel blooms, prioritize at-risk lakes for additional monitoring, and optimize management to protect human health and other environmental end goals.

Plain Language Summary When lakes receive large amounts of nutrients from the surrounding landscape due to fertilizer runoff or other sources of nutrient pollution, they can develop algal blooms. Algal blooms are harmful to the lake ecosystem and sometimes produce toxins which are dangerous to humans and animals. To assess this issue, lake chlorophyll-*a*, a measure of algal presence, is monitored. This monitoring is limited in reach due to the expense of in-lake sampling and the limited resolution of satellite technology. However, there is a wealth of climate, nutrient, landscape, and in-lake characteristic data for the conterminous US (CONUS) which explains much of what contributes to nutrient pollution and algal growth. Here, we use this data in a machine learning model to predict nutrient (total nitrogen and total phosphorus) and chlorophyll-*a* concentrations in about 112,000 lakes in the CONUS. We found that high chlorophyll-*a* concentrations are more likely in the Midwest where agriculture is prevalent, but other areas with high lake chlorophyll-*a* concentrations are present across the CONUS in nutrient pollution hot spots. These predictions of lake nutrient and chlorophyll-*a* concentrations can help managers identify areas of concern, prioritize at-risk lakes for testing, and target management to protect human health and the environment.

1. Introduction

Widespread nitrogen and phosphorus nutrient pollution of lakes, rivers, and streams contribute to eutrophication and the formation of harmful algal blooms (HABs) across the globe. These impacts detract from the ecological services freshwater systems provide (Burford et al., 2018; Compton et al., 2011). Here, HABs are defined as an

algal bloom that has any type of harmful impact including the production of unsightly scum, taste and odor issues, toxins, hypoxia, and fish kills (Gorney et al., 2023; U.S. Environmental Protection Agency, 2023). In lakes, many factors contribute to algal bloom formation, but they tend to more often occur in lake systems with high nitrogen and phosphorus concentrations that fuel increased algal growth (Heisler et al., 2008; Iiames et al., 2021). This increased algal growth, especially under eutrophic or hypereutrophic conditions, contributes to hypoxic and anoxic conditions that compromise fisheries, and in some extreme cases cause fish kills (Watson et al., 2016; Yuan & Pollard, 2015). HABs also diminish the esthetic appeal of aquatic ecosystems compromising the recreational value of the waters (Suplee et al., 2009). In more extreme cases, some HABs generated by cyanobacteria produce cyanotoxins that can harm humans and animals (Burford et al., 2018; Paerl et al., 2001), and here we refer to these more specifically as cyanoHABs. Exposure to cyanotoxins from cyanoHABs through drinking water, fish consumption, and recreation can contribute to numerous health issues, particularly in the liver, kidneys, gut, and respiratory systems, and may be particularly dangerous to those with pre-existing conditions (Chorus & Welker, 2021; Lad et al., 2022). While HABs are generally tied to eutrophic conditions stemming from point and nonpoint sources of nutrient pollution (Heisler et al., 2008), the relative importance of nutrient pollution compared to known modifying factors like in-lake characteristics and other environmental drivers varies based on the scale of the analysis (Glibert, Beusen, et al., 2018; Iiames et al., 2021; Sabo et al., 2023). According to the US National Lakes Assessment in 2012, over 35% of lakes have excess nitrogen and phosphorus and nearly 50% of lakes exceed recreational benchmarks for chlorophyll-*a* (U.S. Environmental Protection Agency, 2016a). To better understand where these lakes are located and how point and nonpoint nutrient sources combine with other environmental factors to produce this finding, we can leverage large scale, nationally consistent nutrient input data.

Eutrophication and HABs are a concern for many lakes across the US. There are numerous approaches for quantifying HABs for lake management and limnological research questions. A number of variables can be measured to quantify the trophic status and magnitude of a bloom including chemical indicators such as nitrogen and phosphorus concentrations and biological indicators such as chlorophyll-*a* and cyanobacteria cell counts. Lake HAB monitoring is especially challenging given the often ephemeral and patchy behavior of blooms (Stumpf et al., 2016). Field assessments and in situ monitoring can determine the extent of HABs and water quality monitoring helps to determine when environmental conditions are likely to lead to HABs occurrences (Glibert, Pitcher, et al., 2018; Kim et al., 2021). However, these efforts are resource intensive and are usually limited to those lakes of interest which warrant the expense (Brooks et al., 2016; Glibert, Pitcher, et al., 2018). In recent years, satellite technologies have been implemented to determine the extent of algal blooms in lakes (Handler et al., 2023; Iiames et al., 2021; Meyer et al., 2024; Naghdi et al., 2020; Shi et al., 2017; Topp et al., 2021). These remote sensing platforms have enabled bloom monitoring for many thousands of waterbodies at monthly to near-daily timescales. Still, remote sensing techniques are limited to the spatial resolution of the satellite with many platforms excluding smaller water bodies (Glibert, Pitcher, et al., 2018; Liu et al., 2022). While remote sensing data has greatly increased the amount of bloom monitoring data, there is rarely concomitant lake nutrient data—that is still largely limited to field monitoring. Understanding where nutrient inputs are contributing to excess nitrogen and phosphorus in lakes can inform which lakes are at higher risk for developing algal blooms. Taking this approach allows for making predictions for a much larger array of lakes across the US, including smaller waterbodies and headwater catchments.

In addition to field monitoring and remote sensing, empirical modeling is a third complementary approach to these efforts which utilizes standardized data sets and a variety of modeling techniques to predict HABs risk in lakes at a variety of scales, including across the CONUS. Nutrients, and the land uses which are associated with high nutrient loads, are by far the most often cited contributor to HABs production (Burford et al., 2018; Butcher et al., 2023; Iiames et al., 2021; Marion et al., 2017), followed by climatic factors (Ho & Michalak, 2020) and watershed characteristics (Iiames et al., 2021). Prior modeling efforts which link national nutrient mass balance data to observed lake and stream nutrient concentrations have highlighted the importance that climate, edaphic, and other ecosystem characteristics have in the retention or loss of nutrients to lacustrine systems (Lin et al., 2021; Sabo et al., 2019; Sabo, Clark, Gibbs, et al., 2021; Sabo et al., 2023; Seegers et al., 2021). In turn, these data sets and nutrient prediction models provide extensive information on the importance of environmental conditions in predicting algal bloom production since they are highly dependent on elevated nutrient concentrations. In this study, we expand upon past modeling efforts by leveraging large spatial data sets that incorporate national nutrient inventories and other environmental indicators to make extensive and fine-scale predictions of

chlorophyll-*a* concentrations. Chlorophyll-*a* concentration serves as a common indicator for HABs risk for jurisdictions across the United States and the globe as it highlights potential risk for emergent hypoxic/anoxic conditions as well as the increased probability of cyanobacterial blooms (Beaver et al., 2018; Chorus & Welker, 2021; Loftin et al., 2016; Yuan et al., 2014; Yuan & Pollard, 2015, 2019). We acknowledge that although chlorophyll-*a* is indicative of algal bloom production, it is not always indicative of toxin-producing cyanoHABs and that there has been varying success in relating chlorophyll-*a* to toxic algal bloom production (Hollister & Kreakie, 2016; Søndergaard et al., 2011; Yuan et al., 2014). In this study, chlorophyll-*a* is used as an indicator for trophic status and algal bloom severity, but not specifically cyanobacteria or cyanotoxin production.

By integrating publicly available, large spatiotemporal data sets within a machine learning framework, we address knowledge gaps in current US nutrient and chlorophyll-*a* monitoring and modeling capabilities. We did this by associating the magnitude of pollution sources and other environmental factors with total nitrogen (TN), total phosphorus (TP), and chlorophyll-*a* concentrations for >2,000 lake observations across the contiguous United States (CONUS) via random forest (RF) (Breiman, 2001; Lin et al., 2021; Sabo et al., 2023). The RF model was then applied to predict the nutrient and chlorophyll-*a* concentrations for ∼112,000 existing in-network lakes as well as hypothetical headwater lake conditions for ∼2.5 million catchments across the CONUS, based on the National Hydrography Data set (NHD+ v2) designations (Geological Survey, 2004). The vast majority of these lakes are not monitored by field crews or satellites due to the inherent limitations of current quantification methods. Hypothetical headwater lake catchment predictions allow for risk prediction maps that can help local jurisdictions and landholders assess likely trophic condition and HABs risk for lakes within their locality that are not in network (lakes that are geographically isolated from the NHD+ stream network, e.g., larger ponds, perched lakes, headwater lakes, and smaller reservoirs). To make predictions, we synthesized spatial data sets in a CONUS-wide analysis and employed a unique nested modeling strategy wherein RF predictions of TN and TP were used to predict chlorophyll-*a*. We assessed the utility of these national data sets in making chlorophyll-*a* predictions across large spatial scales. We explored the mechanisms which lead to increases in chlorophyll-*a* concentrations and the probability of HABs. Finally, we investigated how seasonality and lake depth affect chlorophyll-*a* concentrations predicted by the RF models.

2. Methods

We used lake and watershed characteristic variables as well as time varying climatic and nutrient mass balance predictors for 2,062 U.S. Environmental Protection Agency (EPA) National Lakes Assessment (NLA) lake observations to train RF models for TN, TP, and chlorophyll-*a*. We then applied these models to make predictions of lake TN, TP, and chlorophyll-*a* concentrations for ∼2.5 million US catchments and ∼112,000 US lakes (Figure S1 in Supporting Information S1). Using 2017 NLA thresholds for chlorophyll, we also estimate likely trophic status and algal bloom severity for these lakes.

2.1. Training Data

The data set used to train our RF models was developed in recently published works described in Sabo et al. (2023) and Lin et al. (2021) and contains data from several sources. Response variables of chlorophyll-*a*, TN, and TP (μg/L) are from publicly available NLA surveys for the years 2007 (n = 1,098 samples) and 2012 (n = 964 samples) (U.S. Environmental Protection Agency, 2010, 2016b). Each NLA selects lakes according to a probability-based survey design to be representative of the population of US lakes. The population is defined as lakes that are at least 4 ha in surface area for the 2007 NLA and 1 ha in 2012, in addition to having a depth of at least 1 m. Approximately one third (n = 337) of lakes sampled in 2007 were included in the 2012 NLA. In total, the data set includes 2,062 observations from 1,725 unique lakes. At each lake, a depth-integrated photic zone (maximum depth of 2 m) water sample was collected for water chemistry and chlorophyll-*a* (U.S. Environmental Protection Agency, 2007, 2011, 2012). Samples are collected from the deepest point in the lake and generally exclude shorelines. The depth of the lake is measured at this collection point. Chlorophyll-*a* samples were filtered in the field immediately after collection. Chlorophyll-*a* was extracted with 90% acetone and analyzed by fluorometry. Unfiltered water was subject to persulfate digestion before analyzed for TN and TP concentrations. This data set encompasses a wide spatial range of lake locations across the CONUS and represents a large distribution of observed chlorophyll-*a* concentrations (Figure 1).

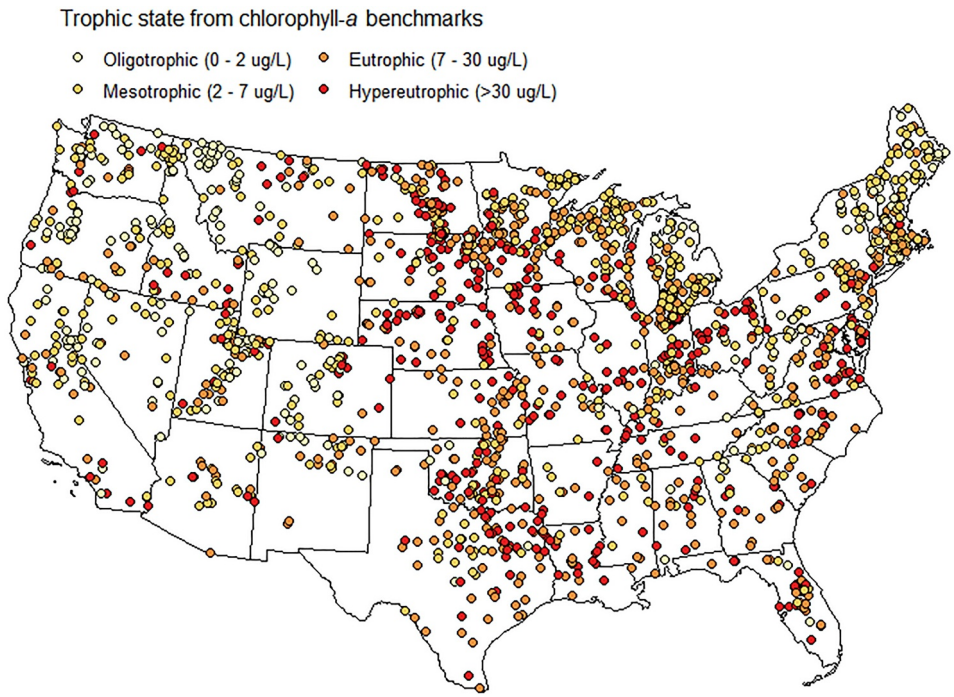

Figure 1. Map of the conterminous United States showing 2007 and 2012 National Lakes Assessment (NLA) observations (n = 2,062) used in training Random Forest models. In cases where there are multiple samples for a lake, the most recent sample value is displayed. Point colors represent chlorophyll-*a* concentrations binned by 2017 NLA trophic state benchmarks.

The predictors in our training data set include lake depth, watershed characteristics (e.g., runoff, erodibility, soil clay content), climatic, ecosystem, and nutrient mass balance variables; all except for lake-specific variables were summarized at the lake watershed scale. Lake depth data are from the same NLA data sets from which the response variables came and were included in our analysis due to the importance of lake depth in explaining spatiotemporal variation in lake nutrient concentrations (Sabo et al., 2023). Watershed characteristic data are from U.S. EPA's LakeCat (Hill et al., 2018). We chose to include LakeCat variables representing land use, edaphic, and hydrologic factors in our training data set due to their importance in the landscape-to-lake nutrient pathway. This data is not temporally specific. Net primary productivity (NPP) data and climatic variables, such as snow cover, precipitation, and land surface temperature, are from PRISM and EON (NASA Earth Observatory Network, n.d., PRISM Climate Group, Oregon State University, n.d.). Finally, we incorporated annually summarized nutrient pollution source variables into the training data set: atmospheric nitrogen, sulfur (National Atmospheric Deposition Program, 2020), and phosphorus (Wang et al., 2017) deposition and nitrogen and phosphorus mass balance estimates from nutrient inventories for the years 2007 and 2012 (Sabo et al., 2019; Sabo, Clark, & Compton 2021; Sabo, Clark, Gibbs, et al., 2021). All data were spatially matched by NLA site ID using the National Aquatic Resources Surveys (NARS) watershed delineations. For temporally-specific variables, data were temporally matched by year and, when possible, month.

Our initial training data set included 52 predictor variables (including observed TN and TP, also used as responses; see Tables S1, S2, and S3 in Supporting Information S1 for a full list of initial predictors) but through a variable selection process, we reduced the number of predictors in our final RF models to 25 (Table 1). Out of the initial training data set, three sets of predictor variable data sets were formed: one data set for training the TN RF which eliminated phosphorus-specific terms (n = 38), one data set for training the TP RF which eliminated nitrogen-specific terms (n = 38), and a final data set which utilized all available variables to train the chlorophyll-*a* RF (n = 52). The predictor variables from each set were run through a principal components analysis (PCA) which determined the variability in the data that each predictor explained (Tables S1, S2, and S3 in Supporting Information S1). We utilized this information, along with expert knowledge about the relative importance of each variable based on previous studies (Lin et al., 2021; Sabo et al., 2023), to remove repetitive predictors. Further

Table 1
Summary of Response and Predictor Variables Used in Training Final Random Forest Models

Variable	Data source	Units	Temporal scale
Responses			
Chlorophyll-*a*	NLA	µg/L	Single day sample
Observed Nutrients Responses and Predictors			
Total nitrogen (TN)	NLA	µg/L	Single day sample
Total phosphorus (TP)	NLA	µg/L	Single day sample
Climate & Ecosystem Predictors			
Annual net primary production (NPP)	EON	g C/d	Monthly - average of previous 12 months
Monthly net primary production (NPP)	EON	g C/d	Monthly
Annual precipitation	PRISM	mm/yr	Monthly - average of previous 12 months
Annual snow cover	PRISM	%	Monthly - average of previous 12 months
Annual temperature	PRISM	°C	Monthly - average of previous 12 months
Monthly temperature	PRISM	°C	Monthly
Nutrient Inventory Predictors			
Agricultural N fertilizer	NNI	kg/ha/yr	Annual
N-fixing crop cultivation	NNI	kg/ha/yr	Annual
Net anthropogenic N input	NNI	kg/ha/yr	Annual
Atmospheric N deposition	NADP	kg/ha/yr	Annual
Atmospheric P deposition	NNI	kg/ha/yr	Annual
Atmospheric S deposition	NADP	kg/ha/yr	Annual
Total P input	NNI	kg/ha/yr	Annual
Accumulated agricultural P input	NNI	kg/ha/yr	Annual
Watershed & Lake Characteristics Predictors			
Lake depth	NLA	m	Single day sample
Median runoff	LakeCat	mm	Static
Agricultural erodibility	LakeCat	NA - factor	Static
Base flow index	LakeCat	%	Static
Clay content	LakeCat	%	Static
Sand content	LakeCat	%	Static
P2O5 content[a]	LakeCat	%	Static
Wetland cover	LakeCat	%	Static
Watershed area	LakeCat	km^2	Static

Note. Does not include variables that were excluded by the variable selection process. [a]P2O5 content refers to the "mean % of lithological phosphorus oxide (P2O5) content in surface or near surface geology" (Hill et al., 2018).

variable selection was achieved by first running each RF model with all remaining predictors which allowed us to determine variable importance for each model predictor. We could then run a series of models starting with just the two most important predictors and iterating through all variables in order of importance. We calculated the R-squared and root mean squared error (RMSE) for each model run. Final models were those with the least number of variables wherein model performance metrics no longer showed marked improvements with any additional predictors (Figures S2, S3, and S4 in Supporting Information S1). It should be emphasized that the elimination of predictors does not necessarily dismiss the importance of removed predictors, but that the duplicative statistical information they provide would offer no improvement to model performance and would only complicate model interpretation. It is important to carefully interpret the results of RF models under the reality of multi-collinearity (Sabo et al., 2023).

2.2. Training Random Forest Models

RF modeling (Breiman, 2001) is a machine learning technique that uses a number of decision trees, each calibrated with a random subset of data, to predict responses. We used RF models because they can manage data sets with many predictors (Cutler et al., 2007). In recent years, RF modeling has become a common tool for making large-scale environmental inferences (Hill et al., 2017; Iiames et al., 2021; Lin et al., 2021; Pennino et al., 2020).

We developed RF models for three responses: TN, TP, and chlorophyll-*a*. For predictions, we used a nested framework wherein predicted TN and TP concentrations from the RF models were inputs for predicting chlorophyll-*a* concentrations. All random forest models were developed in the R statistical software (R Development Core Team, 2022) using the *ranger* package (Wright & Ziegler, 2017). Each RF used 500 trees and the default "mtry" value from *ranger*; we tested a range of values for these settings but found that they did not significantly alter model output. For each RF model, the response variable was transformed with the function log $(x + 1)$ to normalize the distribution and optimize model fitting and prediction (De'ath & Fabricius, 2000; Walsh et al., 2017). Models were trained with 80% of the samples from the training data set and tested with the remaining 20%. As such, each RF model run never considered ∼400 of the lake observations during calibration. In addition, we performed stratified sampling of our training and testing splits, which samples according to the binned distribution of values of the respective responses to ensure that the full spread of the data were represented in each model run. RF models were run 10 times, each time on a different split of the data to allow for cross-validation of modeling results.

After visual inspection of CONUS wide residual maps, we detected no clear evidence of regional bias for the final TN, TP, and chlorophyll-*a* RF models (Figure S5). Relatedly, we attempted regional models and models based on lake classes resulting from a cluster analysis of CONUS lake characteristics and patterns of chlorophyll-*a* responses to nutrients, but these attempts yielded no considerable improvements in model performance, indicating that national models were sufficient (results not included).

To evaluate the success of our modeling efforts, we computed several model performance metrics: the *r*-squared of testing set predictions versus observations, the r-squared of training set predictions versus observations, the RMSE, mean bias, and variance (Pennino et al., 2020). For each RF model, performance metrics were calculated for each of the 10 cross-validation model runs and then averaged. The same was done for variable importance scores. We also assessed model results by looking at partial dependence plots (PDPs) which show the marginal effect of a single predictor variable on the response with the predictor variable on the *x*-axis and the response variable on the *y*-axis. PDPs have a large margin of error but can show the general direction of the relationship between a response and predictor. Unlike with performance metrics, we plotted PDPs individually for each of the 10 model runs and showed only the PDP from the first model run in our results as none were significantly different.

2.3. CONUS-Wide Data for Making Spatial Predictions

We used two different types of data sets to make predictions: one at the NHD+ catchment scale (3 km^2 average) and one at the lake watershed scale (359 km^2 average). Predictions at the catchment scale facilitate making fine-scale, comprehensive assessments of likely chlorophyll-*a* concentrations for NHD+ defined catchments within the CONUS which are applicable to many headwater lakes, ponds, and reservoirs. However, this useful screening map lacks specificity to existing lakes, which necessitated the development of flow accumulated, watershed average estimates of our predictors for ∼112,000 lakes across the CONUS. Predictor variables described in the *Training data* section above were assembled into our catchment and lake watershed prediction data sets.

For the catchment-scale data set, variables were summarized for ∼2.5 million catchments across the entire CONUS regardless of lake presence. Since lake depth data does not exist for every catchment, we tested the effects of theoretical lakes with varying maximum depths on chlorophyll-*a* concentrations by making catchment-scale predictions three times, each with one static maximum lake depth value across the CONUS: a shallow lake (1 m), a mid-sized lake (10 m), and a deep lake (50 m). These values largely capture the distribution of observed and modeled lake depths in LAGOS and NHD+, respectively (Table S4 in Supporting Information S1). Watershed characteristic data (Table 1) are readily available at the catchment scale as StreamCat is a catchment-level complementary data set to LakeCat's lake watershed features (Hill et al., 2018). For climatic and deposition variables, catchment-specific values of desired monthly and annual averages were computed from rasters using

zonal statistics (Figure S6b in Supporting Information S1). Nutrient inventory terms were downscaled from the HUC8 (hydrologic unit boundaries delineate truncated portions of a watershed—HUC8s are equivalent to medium-sized river basins) to the catchment scale by allocating nutrient terms based on land use (essentially, attributing nutrient amounts based on percent of agricultural or urban land in each catchment; Figure S6A; Equations S1 and S2 in Supporting Information S1).

For lake watershed-scale prediction data sets, data was summarized for about 112,000 US lakes and their contributing watersheds. This contrasts with the catchment-scale data sets which present data for individual NHD+ catchments covering the CONUS regardless of lake presence. These lake watershed-scale predictions were instead made for existing lakes and therefore we could utilize lake-specific information when it was available. Lake depth data was obtained from LAGOS where possible (25% of modeled lakes) (Smith et al., 2021) and, if not, was obtained from NHD+ which includes modeled lake depths (Geological Survey, 2004; Hollister et al., 2011). Landscape characteristic data for lake watershed-scale prediction data sets was available through LakeCat. For all other climatic and nutrient mass balance variables, catchment prediction data were summarized at the lake watershed scale through a flow accumulation method which averages across catchments that contribute to a lake's watershed.

2.4. Predictive Modeling

We first used the TN and TP RFs to make predictions of TN and TP concentrations at both the catchment and lake watershed scales. These predicted nutrient values were then used to make predictions of chlorophyll-*a* at the catchment and lake watershed scales by being fed into the chlorophyll-*a* RF in place of observed TN and TP concentrations that were used to train the model. This nested approach was utilized because including TN and TP values in our chlorophyll-*a* RF improved our results markedly and because observed TN and TP values, although available at the scale at which we trained the models (NLA lake watersheds), were not available at the scales at which we made predictions (CONUS-wide catchments and in-network lake watersheds).

We performed predictive modeling for several scenarios. First, as described in the *CONUS-wide data for making spatial predictions* section above, since lake depth data does not exist for every catchment across the CONUS, all catchment-level predictive modeling efforts were performed over a range of theoretical maximum lake depths. Second, to observe the effects of seasonality on chlorophyll-*a* concentrations, we used seasonally-specific values for those climate variables in our final models which are monthly averages—land surface temperature and NPP. We made predictions of TN, TP, and chlorophyll-*a* for May, July, and October of 2007 and 2012 at the catchment scale, representing potential chlorophyll-*a* concentrations for the entire CONUS regardless of lake presence, and at the lake watershed scale, representing lake-specific chlorophyll-*a* concentrations in a portion of existing in-network CONUS lakes. The data set used in training our RF models contains data at the lake watershed scale, so while lake watershed predictions were made at the same scale represented in the training data set, catchment-scale predictions are made at a different scale. These catchment-scale predictions may or may not be representative of off-network, headwater lake watersheds depending on differences in their size and characteristics. As each of our RF models include cross-validation for which 10 model runs were executed, our final predictions averaged results from the 10 runs. For improved visualization and interpretation of results, we used the 2017 NLA trophic state chlorophyll-*a* benchmarks: Oligotrophic = 0–2 μg/L, Mesotrophic = 2–7 μg/L, Eutrophic = 7–30 μg/L, and Hypereutrophic = >30 μg/L. In addition, for catchment-level predictions, we improved visualization by interpolating results across catchments with missing data (4.3%).

3. Results

3.1. Final RF Models

The final TN RF model explains 65% of the variation in TN concentrations for lakes across the cross-validated testing data sets (Table S5). The model has a RMSE of 0.27, a mean bias of −0.00028, and a variance of 0.27 (Table S5 in Supporting Information S1). The TN RF model contains 17 predictor variables: four nutrient inventory variables, six climate & ecosystem variables, and seven watershed & lake characteristic variables (Figure 2a). Lake depth is the top predictor for TN by a large margin and the PDP shows a negative relationship between lake depth and TN concentrations (Figures 2a and 3 [TN-1]). The second most important predictor of TN is agricultural nitrogen fertilizer which has a positive relationship with TN according to the PDP (Figures 2a and 3 [TN-2]). The third and fourth most important predictors in the model are annual NPP and median runoff,

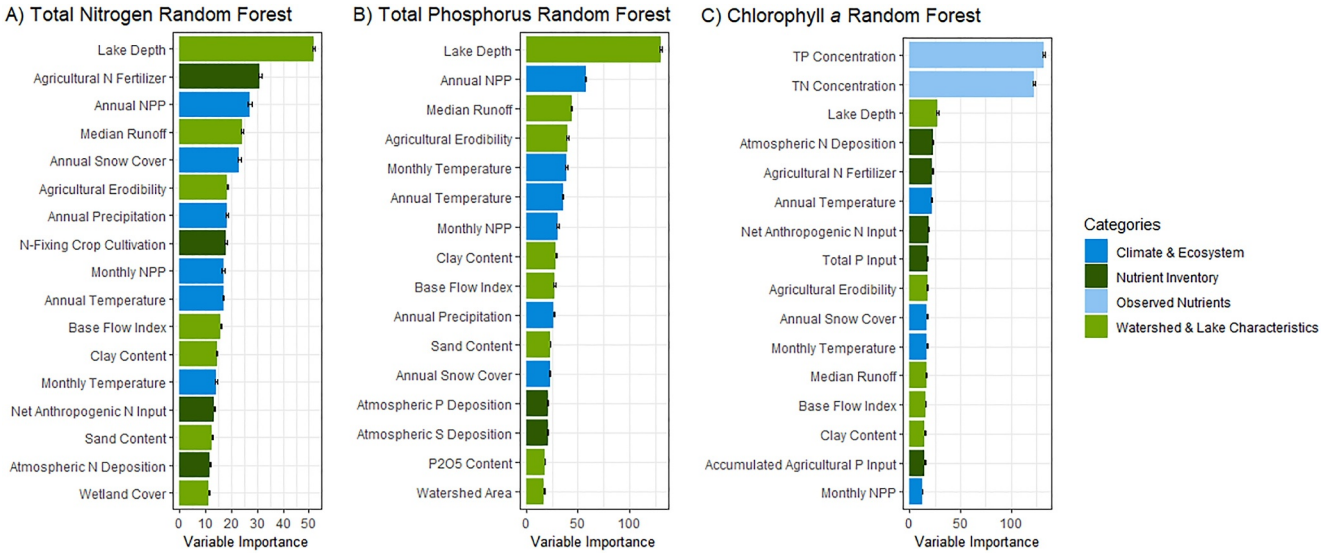

Figure 2. Variable importance rankings for all variables in the final versions of the (a) Total Nitrogen (TN) Random Forest (RF) model, (b) Total Phosphorus (TP) RF model, and (c) Chlorophyll-*a* RF model. The TN model explained 94% of variance in the training data set and 65% in the validation data set, the TP model explained 93% of variance in the training data set and 62% in the validation data set, and the Chlorophyll-*a* model explained 94% of variance in the training data set and 68% in the validation data set (Table S5 in Supporting Information S1) Bar colors indicate variable category. Error bars refer to the standard error calculated among the 10 cross-validation runs for each model.

respectively, both of which have a negative relationship with TN according to PDPs (Figures 2a and 3 [TN-3], and Figure 3 [TN-4]).

The final TP RF model explains 62% of the variation in TP concentrations for lakes across the cross-validated testing data sets (Table S5 in Supporting Information S1). The model has a RMSE of 0.37, a mean bias of

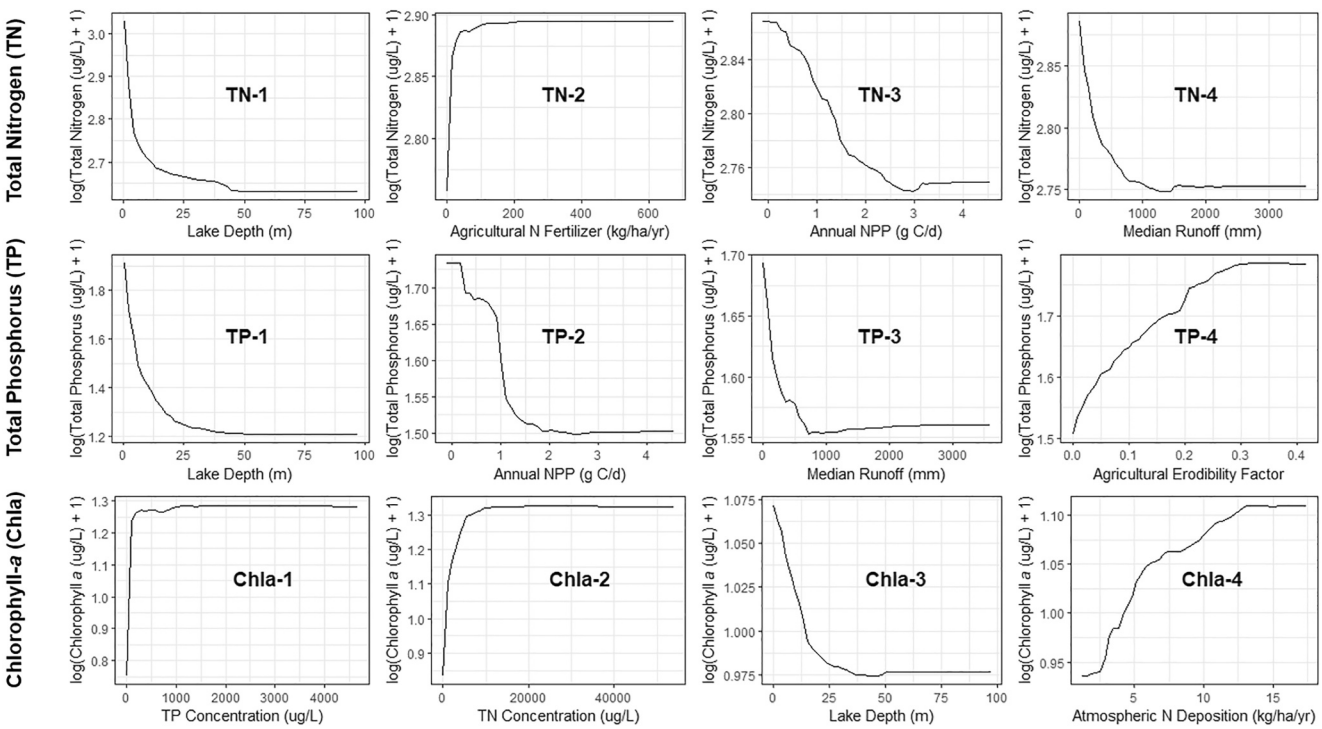

Figure 3. Random Forest (RF) partial dependence plots showing the relationship between the top 4 important predictor variables (panel numbers 1–4 refer to importance ranking) and responses from each RF: Total Nitrogen (TN), Total Phosphorus (TP), and Chlorophyll-*a* (Chla).

0.0034, and a variance of 0.37 (Table S5 in Supporting Information S1). The TP RF model contains 16 predictor variables: two nutrient inventory variables, six climate & ecosystem variables, and eight watershed & lake characteristic variables (Figure 2b). As in the TN RF model, lake depth is the most important predictor for TP by a large margin and the PDP shows a negative relationship between lake depth and TP concentrations (Figures 2b and 3 [TP-1]). The second and third most important predictors in the TP RF are annual NPP and median runoff, respectively (Figure 2b). As in the TN RF model, both annual NPP and median runoff PDPs show a negative relationship with the response variable, in this case TP concentrations (Figure 3 [TP-2] and Figure 3 [TP-3]). The fourth most important predictor of TP is agricultural erodibility which has a positive relationship with TP according to the PDP (Figures 2b and 3 [TP-4]).

The final chlorophyll-*a* RF model explains 68% of the variation in chlorophyll-*a* concentrations for lakes across the cross-validated testing data sets (Table S5). The model has a RMSE of 0.32, a mean bias of −0.0018, and variance of 0.32 (Table S5 in Supporting Information S1). The chlorophyll-*a* RF model contains 16 predictor variables: two observed nutrient variables, five nutrient inventory variables, four climate & ecosystem variables, and five watershed & lake characteristic variables (Figure 2c). TP and TN are by far the most important predictors for explaining the spatiotemporal variation in chlorophyll-*a*, both with a positive relationship with chlorophyll-*a* concentrations according to PDPs (Figures 2c and 3 [Chla-1], and Figure 3 [Chla-2]). Note that variables that are directly influential on TP and TN concentrations in the nutrient RF models ipso facto have a large influence on chlorophyll-*a* concentrations. The third most important predictor in the chlorophyll-*a* RF is lake depth (Figure 2c). As with TN and TP, the PDP shows a negative relationship between lake depth and the response variable of chlorophyll-*a* concentrations (Figure 3 [Chla-3]). The fourth most important predictor of chlorophyll-*a* according to our modeling efforts is atmospheric nitrogen deposition which has a positive relationship with chlorophyll-*a* concentrations according to the PDP (Figures 2c and 3 [Chla-4]). RF predictions of chlorophyll-*a* generally correspond to observed values, but the model has a slight tendency to underpredict for the highest chlorophyll-*a* values (Figure S7 in Supporting Information S1). However, these values are far above the chlorophyll-*a* threshold indicating a hypereutrophic state (the most severe category) according to 2017 NLA trophic state benchmarks and therefore are well represented in our interpretation of the results. In our modeling efforts, we included TN:TP ratio as a potential predictor but found that it did not improve our results or add additional interpretable information.

3.2. Predictions

Catchment-level predictions ($n = \sim$2.5 million) indicate the likely TN, TP, and chlorophyll-*a* concentration for a lake in a catchment based on mean landscape, climate, and nutrient conditions for a given month if a lake of 1, 10, or 50-m depth existed in that catchment (Figure 4). Here we discuss July 2007 predictions since other months (May and October) and years (2012) show similar spatial patterns and attenuation with depth (for comprehensive catchment-level chlorophyll-*a* prediction maps, see Figures S8 and S9 in Supporting Information S1). Similarly, we focus on spatial patterns in catchment predictions with a 1-m maximum lake depth here since patterns are similar for all lake depths tested but, for 10-m and 50-m-deep lakes, are less extreme in their designations and therefore less easily interpretable. Also, the median US lake depth, both modeled and observed, is around 1 m so these maps are likely more representative of the most common lake type (Table S4 in Supporting Information S1).

3.2.1. TN and TP

Catchment nutrient concentration predictions decreased overall with increasing theoretical maximum lake depth (Figure 4). For both TN and TP, the largest region with the highest predicted nutrient concentrations (>75th percentile) is the Midwest. Sporadic clusters of high nutrient concentrations (>75th percentile) also occur throughout the CONUS. However, TP shows a unique pattern of a large grouping of high nutrient concentrations centered in the more arid southwestern CONUS which does not appear in maps of predicted TN concentrations.

3.2.2. Chlorophyll-*a*

The number of catchments, for July 2007, categorized as being oligotrophic or mesotrophic, according to 2017 NLA designations, were predicted to increase with lake depth, while the number of catchments predicted to be eutrophic or hypereutrophic consistently decreased with increasing lake depth (Table 2), revealing attenuation of the extent and magnitude of eutrophication and likely algal bloom development with increased lake depth. The

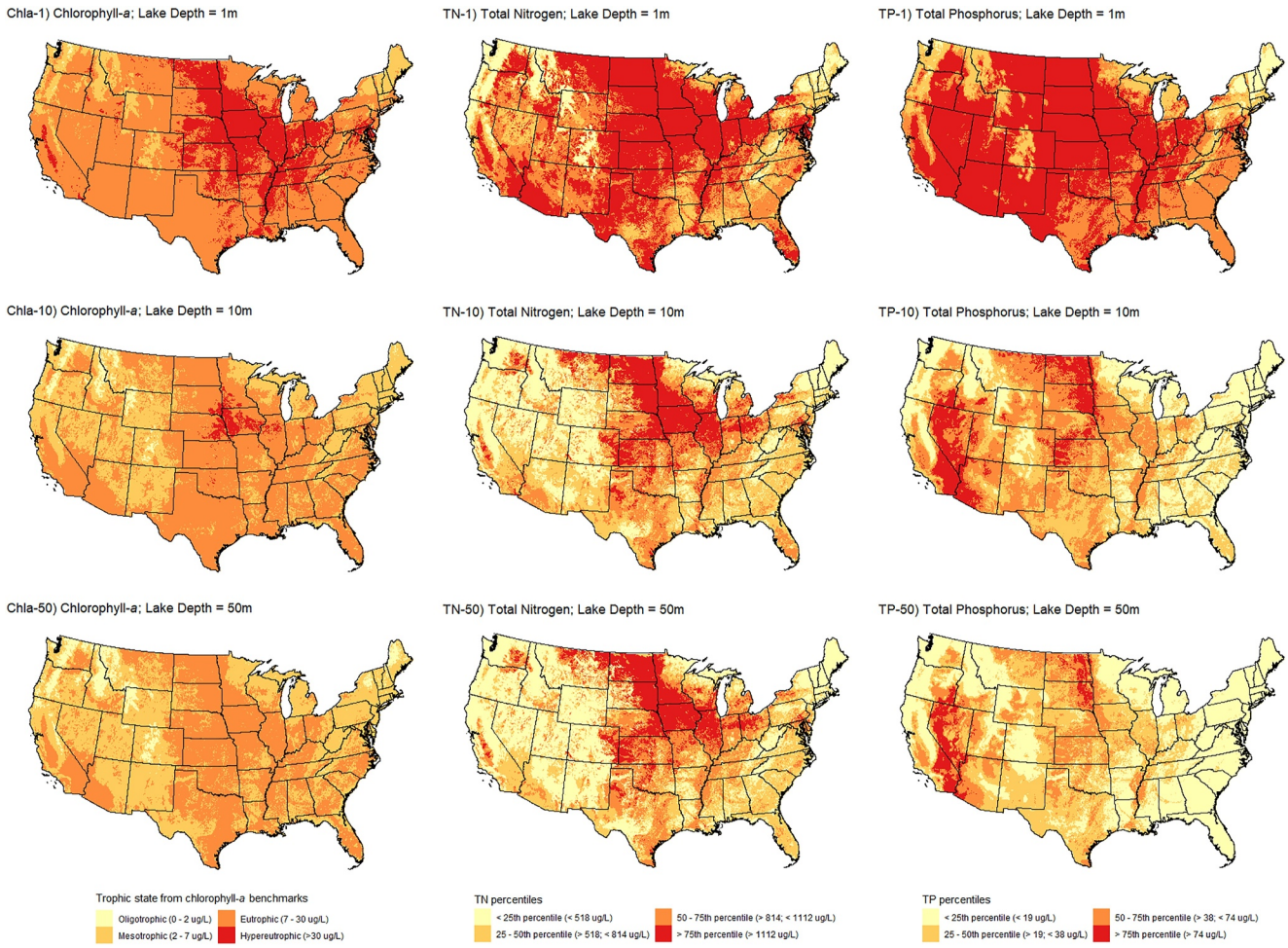

Figure 4. Maps of the conterminous U.S. showing predictions of chlorophyll-*a*, TN, and TP for catchments (n = ~2.5 million) in July of 2007 at three potential maximum lake depths: 1 m, 10 m, and 50 m. For chlorophyll-*a* maps, area colors represent concentrations binned by 2017 NLA trophic state benchmarks. For TN and TP maps, area colors represent concentrations binned by 25th, 50th, and 75th percentiles of combined catchment predictions for July 2007.

region with the most hypereutrophic lakes (>30 µg/L chlorophyll-*a*) according to July 2007 predictions for 1-m-deep lakes is the Midwest with a high-risk area covering portions of many states in the Midwest, central plains, and some east south-central states (North Dakota, eastern South Dakota, Nebraska, Kansas, southern Minnesota, Iowa, Missouri, Illinois, southern Wisconsin, Indiana, Ohio, Kentucky, and Tennessee) (Figure 4a). In addition, portions of this risk area extend down into some southeast and south central states (Oklahoma, Arkansas, eastern Texas, Louisiana, Mississippi, and northern Alabama). Other sporadic clusters of hypereutrophic predictions for 1-m-deep lakes are found near the Great Lakes (southeastern Michigan and western New York state), in the Chesapeake Bay watershed, and in some western states (Idaho and the Central Valley of California).

Table 2
Counts and Percents of July 2007 Catchment Predictions at Maximum Lake Depths of 1, 10, and 50 m in Each Trophic State of the 2017 NLA Chlorophyll-a (Chla) Benchmarks

Lake depth (m)	Oligotrophic (0–2 µg chla L^{-1})		Mesotrophic (2–7 µg chla L^{-1})		Eutrophic (7–30 µg chla L^{-1})		Hypereutrophic (>30 µg chla L^{-1})	
1	2,003	0%	211,936	8%	1,717,159	68%	601,603	24%
10	48,006	2%	1,040,440	41%	1,394,262	55%	49,993	2%
50	127,177	5%	1,373,855	54%	1,031,667	41%	2	0%

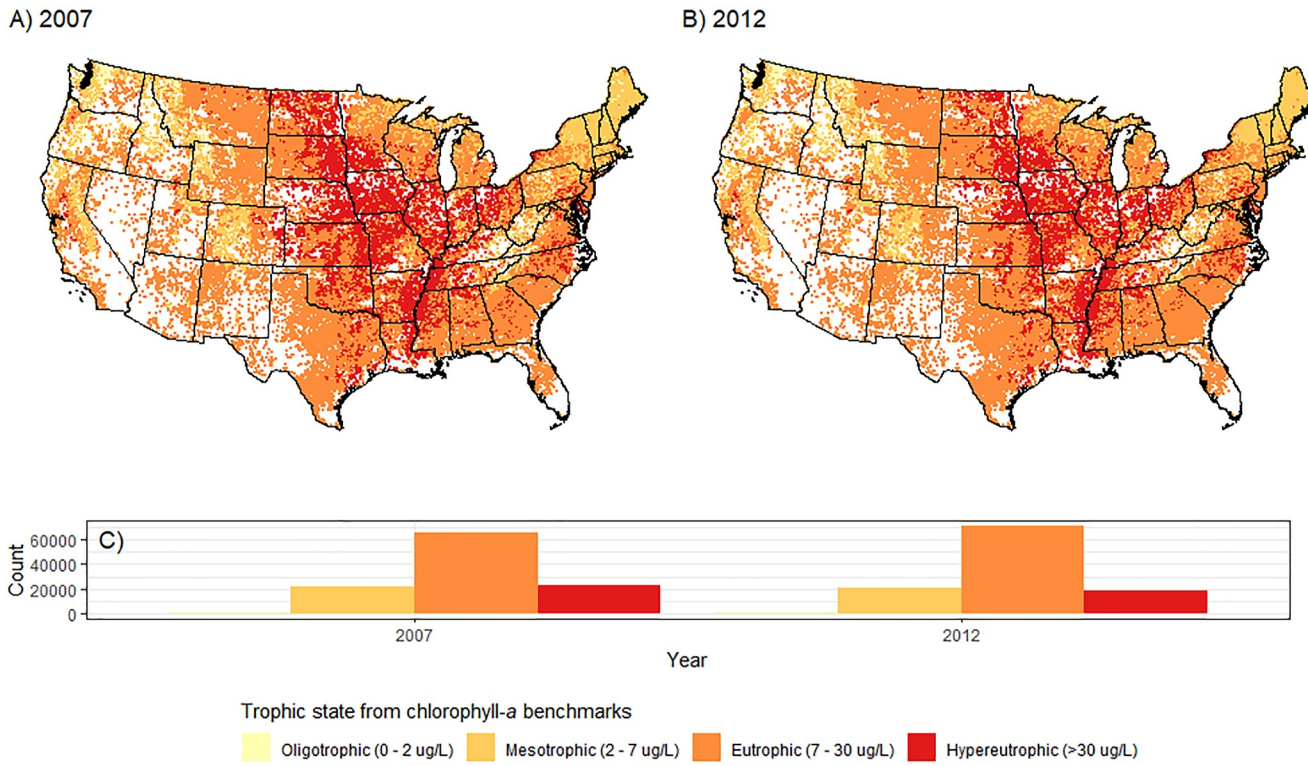

Figure 5. Maps of the conterminous U.S. showing predictions of chlorophyll-*a* for lakes (n = 112,021) in July of (a) 2007 and (b) 2012. Point colors represent chlorophyll-*a* concentrations binned by 2017 NLA trophic state benchmarks. Panel (c) shows counts of these trophic designations for July 2007 and July 2012 lake predictions.

Lake watershed-level predictions of chlorophyll-*a* were made for 112,021 actual lakes in 2007 and 2012 (Figure 5). Unlike with catchment-level predictions, these lake watershed-level predictions are made for existing lakes and incorporate measured and estimated lake depths specific to each lake (described in the *Methods* section above). The spatial pattern of chlorophyll-*a* predictions indicating hypereutrophic lakes is very similar to the catchment-level predictions for 1-m-deep lakes as described above. Also, the spatial pattern of lake watershed chlorophyll-*a* predictions roughly matches between the same month in 2007 (Figure 5a) and 2012 (Figure 5b), although there is some slight difference in the counts of trophic state designations (Figure 5c; for comprehensive lake watershed-level prediction maps, see Figure S10 in Supporting Information S1).

Seasonal patterns in chlorophyll-*a* predictions indicate that from May to July of 2007 there was an increase in hypereutrophic lakes relative to a threshold of 30 µg/L chlorophyll-*a*, particularly in the eastern half of the CONUS (Figure 6a). During this period, changes to above threshold (n = 3,812) outweighed changes to below threshold (n = 1,464). From July to October of 2007 there was a general decrease in hypereutrophic lakes relative to a threshold of 30 µg/L chlorophyll-*a,* although there were some areas such as the Central Valley of California and the center of the CONUS where the number of hypereutrophic lakes increased (Figure 6b). During this period, changes to below threshold (n = 3,790) heavily outweighed changes to above threshold (n = 669; for seasonal change maps for 2012, see Figure S11).

4. Discussion

We used widely available spatial data sets with nutrient, edaphic, landscape, and climate data to make predictions of lake nutrient and chlorophyll-*a* concentrations. The predictions compliment observational data from field monitoring and remote sensing efforts, which quantify nutrient and/or chlorophyll-*a* concentrations in a finite number of lakes. Currently available nutrient and chlorophyll-*a* measurements rely on resource intensive in-person sampling, in situ monitoring (e.g., buoy sensor systems), and the limited resolution of remote sensing satellites (Brooks et al., 2016; Glibert, Pitcher, et al., 2018; Liu et al., 2022). To expand the spatial extent and resolution of information available to understand patterns in nutrients and chlorophyll to aid investigating lake

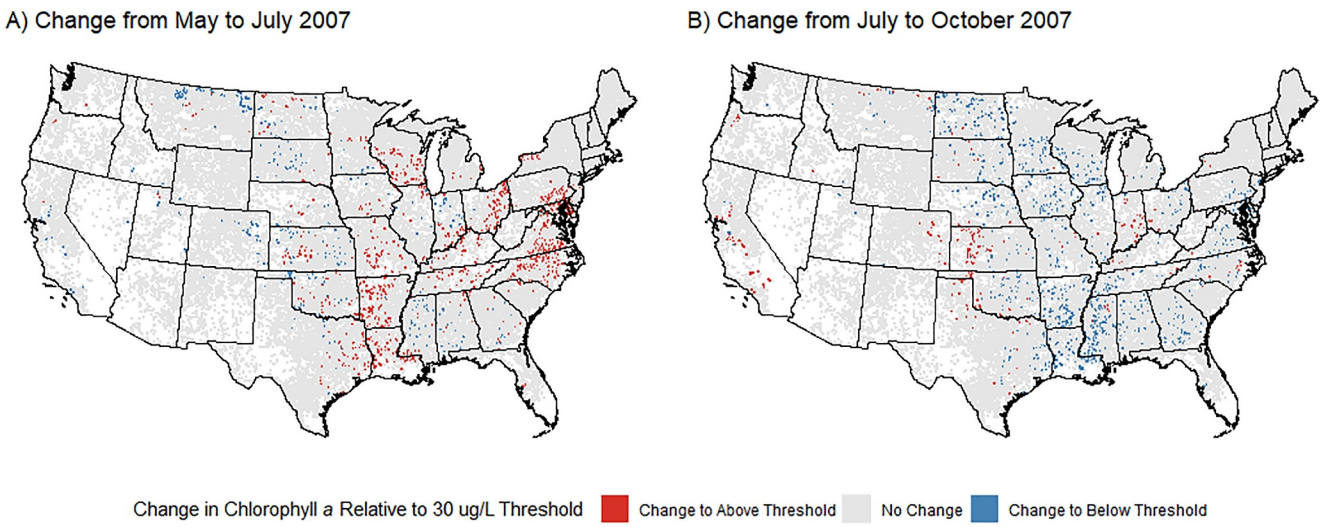

Figure 6. Maps of the conterminous U.S. showing change in predictions of chlorophyll-*a* for lakes (n = 112,021) between (a) May and July and (b) July and October of 2007. Point colors indicate change in chlorophyll-*a* concentration relative to a 30 µg/L threshold which corresponds to a hypereutrophic trophic state according to 2017 NLA benchmarks.

risk for eutrophication, hypoxia, and HABs, we generated predictions for ∼2.5 million catchments and 112,021 lakes across the CONUS. We employed a novel nested modeling approach wherein TN and TP RF predictions were fed into a chlorophyll-*a* RF model resulting in improved predictions of chlorophyll-*a*. Our study implemented machine learning techniques with large data sets to make insights at a broad spatial scale. Notably, using the LakeCat and national nutrient inventory data sets (Hill et al., 2018; Sabo et al., 2019; Sabo, Clark, & Compton 2021; Sabo, Clark, Gibbs, et al., 2021) allowed us to draw explicit connections between landscape characteristics, nutrient pollution, and chlorophyll-*a*. Thus, the results of the RF models can be leveraged to screen for eutrophication, hypoxia, and HABs likelihood in a given area and provide actionable information for management on predominant drivers of nutrient pollution in the lake watersheds of interest.

This work builds on previous modeling efforts that used various combinations of in-lake/stream and landscape level predictors (Brooks et al., 2022; Hollister et al., 2016; Sadayappan et al., 2022; Shen et al., 2020). Specifically, our work expands on such previous modeling efforts by using CONUS-scale variables converted to the resolution of an NHD+ catchment to make TN, TP, and chlorophyll-*a* predictions for a large number of in-network lakes (∼112,000) and potential off-network lakes (all ∼2.5 million NHD+ catchments in the CONUS).

4.1. TN and TP Concentrations in Lakes Are Driven by Inputs, Erosion, and Internal Lake Factors

After accounting for lake depth, we found that the magnitude of agricultural inputs largely drive nutrient gradients across US lakes. The effect of agricultural and other anthropogenic inputs of nutrients on lakes are diminished by environmental contextual variables like lake depth, climate, and watershed NPP rates. Previous work carried out relatively intensive investigations using a similar predictor-response variable data set to elucidate the likely drivers of surface water TN and TP concentrations across the CONUS, and our results were largely consistent with their findings (Lin et al., 2021; Sabo et al., 2023). However, our approach is a valuable step forward in terms of data processing, analysis, and prediction. We developed a standardized, repeatable framework that is computationally efficient to downscale county and HUC-8 level observations to the NHD+ scale to facilitate prediction at ∼2.5 million NHD+ catchments across the CONUS and then leveraged a flow-accumulation procedure to generate 112,021 in-network lake specific predictions across the CONUS. Similar procedures can be deployed to generate stream reach-level predictions of nutrient conditions, and combined, the lake and stream nutrient concentration maps could offer an unprecedented means for decision makers to identify urban and agricultural nutrient pollution hot spots as indicated by the nutrient inventories (Sabo, Clark, & Compton, 2021) and assess expected growing season nutrient concentrations. Various screening exercises can be deployed using these predictions to help prioritize watersheds for greater monitoring and nutrient reduction efforts.

At the CONUS level, the highest predicted nutrient concentrations for lakes occur in major agricultural regions of the country (e.g., Midwest, southeastern PA, California's Central Valley). It also seems aridity plays a major role in the magnitude of nutrient concentrations, although this effect is much more pronounced for TP than for TN. Interestingly, despite the strong attenuating effect of lake depth on nutrients (potential mechanisms discussed in Section 4.2), predicted TN concentrations in much of the Midwest remain high (>75th percentile) even for medium depth (10 m) and deep (50 m) lakes. Much of the warm and humid southeastern US have lower nutrient concentrations that generally fall below the 50th percentile of predicted lake nutrient values. However, many of the predicted chlorophyll-*a* concentrations in this region fall into the eutrophic/hypereutrophic categorization highlighting the potential role of climate and other factors in enhancing algal growth. Overall, linking the magnitude of pollution sources, the weather conditions at the time of sampling, and other environmental factors highlights the diverse suite of drivers on TN, TP, and chlorophyll-*a* concentrations across the US.

4.2. Chlorophyll Is Strongly Driven by Phosphorus and Nitrogen in Lakes Across the US

TN and TP are by far the most important predictors of chlorophyll-*a*. Both TN and TP positively impact chlorophyll-*a* concentrations and thus, increase the probability of algal bloom formation (Figures 2c and 3 [Chla-1], and Figure 3 [Chla-2]). This trend conforms to previous chlorophyll-*a* modeling exercises (Hollister et al., 2016) and our expectations, as chlorophyll-*a* is heavily dependent on nutrient availability. Since these terms are such important predictors of chlorophyll-*a*, we used a nested modeling approach, building on previous work linking the national nutrient inventories with growing season nutrient concentrations (Lin et al., 2021; Sabo et al., 2023). We were initially concerned that the introduction of another level of uncertainty in our TN and TP RFs would compromise our final chlorophyll-*a* predictions. However, in developing our chlorophyll-*a* RF, we experimented with subbing predicted TN and TP into our training data set instead of observations and found that the model performance results were quite similar (R-squared degradation = 0.05) which endorses our modeling technique as reasonably robust to this source of error. Thus, the ability to effectively model nutrient concentrations in lakes across the CONUS (Lin et al., 2021; Sabo et al., 2023), which is similar to or exceeds previous modeling efforts of nutrients and/or chlorophyll-*a* (Brooks et al., 2022; Hollister et al., 2016) allows researchers and managers alike to project likely chlorophyll-*a* concentrations and, by extension, an assessment of surface water eutrophication status, hypoxia, and HABs risk.

While it's common knowledge that surface water TN and TP concentrations are important determinants of lake trophic conditions, their importance relative to each other is difficult to parse out given the collinearity of these nutrient pollution sources and associated predictions of lake nutrient and chlorophyll-*a* concentrations. The relationship between total nutrients and chlorophyll-*a* is further complicated because much of the TN and TP in lakes is contained within chlorophyll-*a* during active blooms (Yuan & Jones, 2020). While TP had higher variable importance in predicting chlorophyll-*a* in our models (Figure 2c), CONUS maps of chlorophyll-*a* catchment-level predictions seem to more closely follow spatial patterns of TN concentration predictions (Figure S12 in Supporting Information S1). Any difference in the importance of TN and TP is likely insignificant because phosphorus and nitrogen pollution are commonly co-occurring, especially across the spatial scales that we examined, and these nutrients are often co-limiting to primary productivity (Burford et al., 2023). In light of the ongoing debate about the relative importance of N versus P in fueling algal growth in lakes (Liang et al., 2020), we added an in-lake concentration TN:TP ratio predictor and recalibrated the RF model. However, this ratio offered little to any improvement in model performance, so it was dropped from further consideration. This null result does not necessarily dismiss the importance of nutrient stoichiometry for algal growth or community composition, but no unique information was offered by it relative to the other terms.

Lake depth is the next most important predictor of chlorophyll-*a* after TN and TP (Figure 2c). Lake depth is also an extremely important predictor of TN and TP in their respective RF models (Figures 2a and 2b), as found in previous work (Sabo et al., 2023). These TN, TP, and chlorophyll-*a* concentrations correspond to near-surface samples in the training data set from the NLA; nevertheless, we can identify and expound on patterns between nutrient and chlorophyll-*a* values and lake depth while acknowledging that we do not have data to account for vertical gradients throughout the water column. PDPs show negative relationships between lake depth and TN, TP, and chlorophyll-*a* (Figure 3 [TN-1], Figure 3 [TP-1], and Figure 3 [Chla-3]). In addition, comparing maps of July 2007 catchment-level chlorophyll-*a* predictions made for the three different potential lake depths reveals that there is a clear attenuation of the extent and magnitude of HABs development with increased lake depth (Figure 4). Deeper lakes tend to have a longer residence time (Brooks et al., 2014) allowing for more internal lake

cycling, P sedimentation (Brett & Benjamin, 2008; Fee, 1979; Sabo et al., 2023), and N processing (Tong et al., 2019) that leads to lower nutrient levels and thus lower chlorophyll-*a* concentrations (Nietch et al., 2022; Wang et al., 2013). Although this is the primary mechanism through which lake depth impacts nutrients and chlorophyll-*a*, there are other mediating factors which may also have an effect.

Lake nutrient and chlorophyll-*a* concentrations are moderated by lake mixing regimes, a complex phenomenon that involves the interaction of climate, salinity, morphology, and depth (Adams et al., 2021; Lewis Jr, 2011). Deeper lakes tend to stratify, whereas shallow lakes tend to be more well-mixed and are therefore not isolated from internal nutrient sources from the sediments and can be more productive (Bonilla et al., 2023; Chen et al., 2018; Ding et al., 2018; Isles et al., 2015). In addition, the relationship between lake depth and lake volume leads to lower constituent concentrations through dilution in lakes with large maximum depths which often facilitate larger water volumes. For chlorophyll-*a*, deeper lakes have a lower proportion of the water column hospitable to growth due to lower light availability and lower water temperatures at greater depths. While phytoplankton may grow where there is plentiful access to light, often near the surface, lake mixing, particularly in shallower lakes, can cause the biomass to be redistributed throughout the water column, contributing to more stable chlorophyll-*a* concentrations throughout the growing season (Kosten et al., 2012; Taranu et al., 2012).

Other top variables for predicting TN and TP are median runoff and monthly NPP (Figures 2a and 2b), both of which have negative relationships with the nutrient response terms (Figure 3 [TN-3], Figure 3 [TN-4], Figure 3 [TP-2], and Figure 3 [TP-3]). Dilution is the mechanism through which median runoff lowers growing season TN and TP concentrations (Kleinman et al., 2006; Zhang, 2018), and can cause nutrient source limitation due to either idiosyncratic hydrologic or biogeochemical factors. Greater NPP, associated with more plant growth and vegetative cover, decreases erodibility on the landscape and encourages more uptake of nutrients, therefore limiting nutrient inputs into lakes and attenuating TN and TP concentrations (Lovett & Goodale, 2011). In areas of the country with higher median runoff, there are higher NPP rates on an annual time scale, thus compounding the effects of nutrient sinks and dilution (Sabo et al., 2023; Zhang, 2018). Other top predictors in our nutrient RFs were agricultural N fertilizer in our TN model (Figure 2a) and agricultural erodibility in our TP model (Figure 2b), both of which have a positive relationship with their respective response variables (Figure 3 [TN-2] and Figure 3 [TP-4]). In examining our variable sets for our three final RF models, we see fewer nutrient inventory terms than expected, however, many of the agricultural variables such as agricultural erodibility are present in the models and explain much of the same variability that the nutrient terms associated with agricultural land use would (Figure 2). As a general trend, we found that variables tied to agriculture drive a positive gradient in nutrient concentrations while other environmental factors such as lake depth, median runoff, and monthly NPP attenuate nutrient loads.

4.3. Chlorophyll Predictions—Spatial and Temporal Patterns

In addition to chlorophyll-*a* predictions for 112,021 real lakes, we also made predictions for ~2.5 million catchments across the US regardless of lake presence (Figure 4). The value of these predictions is that they provide estimates for off-network, headwater lakes that do not fall along the NHD+ stream network and smaller stream impoundments that were not included in the lake-specific predictions that we made. These predictions show a general risk prediction map for all possible locations in the CONUS and more broadly illustrate areas of the country that are at high risk of eutrophication and HABs development based on nutrient, climate, and landscape factors.

Even after accounting for the flow accumulation into lake watersheds, catchment- and lake watershed-scale predictions show spatial patterns that are largely the same (Figures 4 and 5). This highlights that headwater conditions largely propagate to downstream lake systems (Frei et al., 2021). Some aspects of these spatial patterns can be related to known landscape and climate characteristics of the CONUS. It's likely that higher chlorophyll-*a* concentrations predicted in the Midwest and Central Valley of California are related to agricultural production in these regions (U.S. Department of Agriculture, 2021). The relatively higher chlorophyll-*a* concentrations in the southeastern US (Figure 5) may be related to generally higher land surface temperatures in these regions and higher rates of livestock/poultry production (Sabo et al., 2019; Sabo, Clark, & Compton, 2021; Sabo, Clark, Gibbs, et al., 2021; U.S. Department of Agriculture, 2021).

Although many of the variables in our data sets were static or matched annually, we were able to use monthly land surface temperature and watershed NPP data to observe the impact of seasonal shifts on our predictions of

chlorophyll-*a* and their spatial patterns. Due to increased temperatures and shifts in NPP, there is a general increase from May to July as indicated by the greater number of sites with predictions that increased to above a 30 μg/L chlorophyll-*a* threshold as compared to the number of sites that decreased to below the threshold (Figure 6a). The opposite is found for changes from July to October which show a general decrease in chlorophyll-*a* concentrations due to falling temperatures and changes in NPP (Figure 6b). These seasonal patterns in chlorophyll-*a* are consistent with the often-observed pattern of greater lake HABs prevalence in the hottest summer months (Brooks et al., 2017; Coffer et al., 2020). We discerned that the sites with seasonal changes in chlorophyll-*a*, both increasing from the spring to the summer and decreasing from the summer to the fall, are primarily located in the eastern CONUS. This spatial pattern might be attributed to the prevalence of non-arid biomes with widely variable terrestrial NPP values and seasonal shifts in monthly temperature and precipitation (Kicklighter et al., 1999). We note that a small fraction of lakes shows an opposite seasonal shift (decreasing from spring to summer or increasing from summer to fall), generally occurring in states west of the Mississippi River. Since many factors are at play and this model only looked at the impacts of monthly temperature and NPP, it is difficult to parse out the exact effects. Overall, information on seasonal and spatial patterns in chlorophyll-*a* concentrations is highly valuable to managers as it relates to HABs outbreaks, hypoxic events, and eutrophication, and highlights lake sensitivities to shifts in nutrient pollution sources, temperature, precipitation, and other factors.

We executed one last screening exercise to highlight where predicted TN, TP, and chlorophyll-*a* concentrations exceeded the 90th percentile in predicted concentration across the CONUS (Figure S13 in Supporting Information S1). All of these lakes were found in the north-central US (upper Mississippi region of the Dakotas, Minnesota, Iowa, Nebraska, Kansas, Colorado, and Illinois). A large fraction of lakes with these concurrent exceedances falls along or west of the 100th meridian line of the United States. This area is arid but highly agricultural (Sabo et al., 2019; Zhang et al., 2021). This combination leads to difficulty in attaining high nutrient use efficiency in crop production, leading to higher rates of nutrient pollution (Sabo, Clark, & Compton, 2021), and likely causes lakes to be less capable of attenuating heightened nutrient loads relative to other water bodies in wetter climates and different edaphic and other environmental conditions. These lakes represent where optimizing agricultural nutrient management may be most challenging due to multiple factors contributing to poor environmental conditions and HABs risk.

4.4. Limitations

While our models are well-performing and the general regional and seasonal patterns of chlorophyll-*a* concentrations can be inferred from our predictions, we note some important limitations in our work. Throughout this study, we leveraged chlorophyll-*a* as a proxy for HABs—although this is largely accurate for a generalized term of HABs which includes any algal bloom which causes human nuisance and/or ecological harm, this is not entirely appropriate for the narrower definition of HABs which refers to blooms which produce toxins. While many have found an increased likelihood of toxin presence with higher chlorophyll-*a* concentrations, the relationship is not absolute and there are limitations to what can be learned about toxic algal blooms based on chlorophyll-*a* predictions (Hollister & Kreakie, 2016; Pip & Bowman, 2014; Yuan et al., 2014; Yuan & Pollard, 2019). Still, chlorophyll-*a* thresholds are used in the World Health Organization's Alert Level Framework as one entry point to triggering management actions when toxin analysis is not available and to provide further protections to health effects of blooms that are not attributable to cyanotoxins (Chorus & Welker, 2021). To interpret and visualize our chlorophyll-*a* predictions, we utilized the 2017 NLA trophic state chlorophyll-*a* benchmarks. These categories are not authoritative regulatory standards and should not be considered as established chlorophyll-*a* limits but rather a useful example for interpreting our prediction results. The model itself generates numeric estimates of chlorophyll-*a*, TN, and TP, and readers/users can define their own risk categories as needed for their applications.

The performance of our RF models is limited by data availability. Our training data set is made up of data from NLA lakes, which are constrained to being larger than 4 ha in 2007 or larger than 1 ha in 2012 and deeper than 1 m. Hence, our predictions for lakes with a wide range of depths and sizes means that some may lay outside of the prediction space from our training data set. As for maximum lake depth values in our lake watershed-scale prediction data set, for lakes where depth observations from LAGOS were not available, we used modeled lake depths from the NHD+ attributes. It is important to acknowledge that these modeled lake depths contain

some uncertainty (Hollister et al., 2011; Stachelek et al., 2022), however, others have found that using these depth estimates, even with their degree of uncertainty, improves modeling results (Milstead et al., 2013).

We recognize that the data available for these US lakes is substantial, and that other areas may not have the spatial density or amount of data for their lakes. The top 5 predictors of chlorophyll-*a* (TN, TP, lake depth, major inputs, and temperature) explained 66% of the variation. Efforts to apply this approach to other areas might focus on assembling these data from global models or limited local data.

Another limitation in the modeling technique we use is that we only incorporate maximum lake depth and assume that lakes are well mixed (NLA samples the photic zone at the deepest part of the lake). Algal blooms can be patchy within lakes, often accumulating along shorelines due to wind or sources of nutrients (Brookfield et al., 2021; Stumpf et al., 2016). Our models do not capture these HAB problem areas since we treated the lake as a single uniform entity. However, if a user wished to customize the data to predict individual branches of dendritic lakes (e.g., apply multiple lake depths and custom watershed inputs) then the model can easily accommodate this.

While we can make many mechanistic insights from our modeling results, our goals in this study were more oriented towards making predictions. The mechanisms of HABs growth explored in our study are by no means exhaustive and we speculate that some variables in our models could be substituted out with other variables and model performance would be similar. Many of the predictor variables in our spatial data sets were correlated to varying degrees (Lin et al., 2021; Sabo et al., 2023) and those variables that were not included in our final models should not necessarily be dismissed as unimportant. We grant that the combinations of variables in our final models capture much of the effects of other variables that also have important mechanistic pathways for increasing lake nutrient and chlorophyll-*a* concentrations (e.g., legacy P surplus is closely related to agricultural activities represented by other included terms). These additional, correlated variables should be considered when prioritizing watersheds for restoration and crafting nutrient reduction strategies.

4.5. Conclusions

In making broadscale predictions of likely HABs presence for lakes across the CONUS, we identify lakes at risk of ecological degradation (e.g., hypoxia, fish kills) and potential community exposure to toxins through drinking water and recreation. These risk prediction maps can potentially assist regional- and state-level managers in identifying areas of concern from HABs within their jurisdiction and assist in crafting remedies to their nutrient pollution challenges. In addition to its management implications, we see our study as part of a larger trend in science which seeks to use big data and machine learning to shape the understanding of emerging issues and give insight into the potential long-term impacts of anthropogenic disturbances. The future growth and extent of HABs, and its effect on freshwater ecosystems, recreation, and human health, is largely uncertain due to prospective changes in nutrient pollution regimes and climate (Butcher et al., 2023; Ho & Michalak, 2020; Scavia et al., 2021). Increased warming, in particular, may indirectly and directly increase the occurrence of HABs by enhancing in-lake nutrient concentrations and boosting lake metabolism. Subsequent research on HABs risk may be able to use this modeling technique to make predictions under potential future conditions.

Overall, this empirical modeling approach, which integrates large-scale landscape and surface water nutrient and chlorophyll-*a* data, generated robust in-lake predictions of chlorophyll-*a* and nutrients which can be used by lake managers for prioritizing lakes and regions potentially at risk for HABs, to ensure healthy lakes for humans and aquatic life.

Data Availability Statement

The data on which this article is based are available in Brehob et al. (2024).

Acknowledgments
We thank Dr. Christopher Clark and Dr. Lester Yuan who provided helpful comments and review of the manuscript prior to submission, as well as Marc Weber, Dr. Ryan Hill, and Dr. Jeff Hollister for their technical assistance in assembling the data sets used. The National Lakes Assessment 2007 and 2012 data were a result of the collective efforts of dedicated field crews, laboratory staff, data management and quality control staff, analysts and many others from EPA, states, Tribes, federal agencies, universities, and other organizations. This research was made possible in part by an appointment to the Research Participation Program for the US Environmental Protection Agency, Office of Research and Development, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the US Department of Energy and EPA. Any views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

References

Adams, H., Ye, J., Persaud, B., Slowinski, S., Kheyrollah Pour, H., & Van Cappellen, P. (2021). Chlorophyll-a growth rates and related environmental variables in global temperate and cold-temperate lakes. *Earth System Science Data*, *14*(11), 5139–5156. https://doi.org/10.5194/essd-14-5139-2022

Beaver, J. R., Tausz, C. E., Scotese, K. C., Pollard, A. I., & Mitchell, R. M. (2018). Environmental factors influencing the quantitative distribution of microcystin and common potentially toxigenic cyanobacteria in US lakes and reservoirs. *Harmful Algae*, *78*, 118–128. https://doi.org/10.1016/j.hal.2018.08.004

Bonilla, S., Aguilera, A., Aubriot, L., Huszar, V., Almanza, V., Haakonsson, S., et al. (2023). Nutrients and not temperature are the key drivers for cyanobacterial biomass in the Americas. *Harmful Algae*, *121*, 102367. https://doi.org/10.1016/j.hal.2022.102367

Brehob, M. M., Pennino, M. J., Handler, A. M., Compton, J. E., Lee, S. S., & Sabo, R. D. (2024). Estimates of lake nitrogen, phosphorus, and chlorophyll-a concentrations to characterize harmful algal bloom risk across the United States. [Dataset]. *U.S. EPA Office of Research and Development (ORD)*. https://doi.org/10.23719/1529835

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Brett, M. T., & Benjamin, M. M. (2008). A review and reassessment of lake phosphorus retention and the nutrient loading concept. *Freshwater Biology*, *53*(1), 194–211. https://doi.org/10.1111/j.1365-2427.2007.01862.x

Brookfield, A. E., Hansen, A. T., Sullivan, P. L., Czuba, J. A., Kirk, M. F., Li, L., et al. (2021). Predicting algal blooms: Are we overlooking groundwater? *Science of the Total Environment*, *769*, 144442. https://doi.org/10.1016/j.scitotenv.2020.144442

Brooks, B. W., Lazorchak, J. M., Howard, M. D., Johnson, M. V. V., Morton, S. L., Perkins, D. A., et al. (2016). Are harmful algal blooms becoming the greatest inland water quality threat to public health and aquatic ecosystems? *Environmental Toxicology and Chemistry*, *35*(1), 6–13. https://doi.org/10.1002/etc.3220

Brooks, B. W., Lazorchak, J. M., Howard, M. D., Johnson, M.-V. V., Morton, S. L., Perkins, D. A., et al. (2017). In some places, in some cases and at some times, harmful algal blooms are the greatest threat to inland water quality. *Environmental Toxicology and Chemistry*, *36*(5), 1125–1127. https://doi.org/10.1002/etc.3801

Brooks, J. R., Compton, J. E., Lin, J., Herlihy, A., Nahlik, A. M., Rugh, W., & Weber, M. (2022). δ15N of Chironomidae: An index of nitrogen sources and processing within watersheds for national aquatic monitoring programs. *Science of the Total Environment*, *813*, 151867. https://doi.org/10.1016/j.scitotenv.2021.151867

Brooks, J. R., Gibson, J. J., Birks, S. J., Weber, M. H., Rodecap, K. D., & Stoddard, J. L. (2014). Stable isotope estimates of evaporation: Inflow and water residence time for lakes across the United States as a tool for national lake water quality assessments. *Limnology & Oceanography*, *59*(6), 2150–2165. https://doi.org/10.4319/lo.2014.59.6.2150

Burford, M. A., Hamilton, D. P., & Wood, S. A. (2018). Emerging HAB research issues in freshwater environments. Global ecology and oceanography of harmful algal blooms. 381–402.

Burford, M. A., Willis, A., Xiao, M., Prentice, M. J., & Hamilton, D. P. (2023). Understanding the relationship between nutrient availability and freshwater cyanobacterial growth and abundance. *Inland Waters*, *13*(2), 143–152. https://doi.org/10.1080/20442041.2023.2204050

Butcher, J. B., Fernandez, M., Johnson, T. E., Shabani, A., & Lee, S. S. (2023). Geographic analysis of the vulnerability of US lakes to cyanobacterial blooms under future climate. *Earth Interactions*, *27*(1), e230004. https://doi.org/10.1175/ei-d-23-0004.1

Chen, D., Shen, H., Hu, M., Wang, J., Zhang, Y., & Dahlgren, R. A. (2018). Legacy nutrient dynamics at the watershed scale: Principles, modeling, and implications. *Advances in Agronomy*, *149*, 237–313. https://doi.org/10.1016/bs.agron.2018.01.005

Chorus, I., & Welker, M. (2021). *Toxic cyanobacteria in water: A guide to their public health consequences, monitoring and management*. Taylor & Francis.

Coffer, M. M., Schaeffer, B. A., Darling, J. A., Urquhart, E. A., & Salls, W. B. (2020). Quantifying national and regional cyanobacterial occurrence in US lakes using satellite remote sensing. *Ecological Indicators*, *111*, 105976. https://doi.org/10.1016/j.ecolind.2019.105976

Compton, J. E., Harrison, J. A., Dennis, R. L., Greaver, T. L., Hill, B. H., Jordan, S. J., et al. (2011). Ecosystem services altered by human changes in the nitrogen cycle: A new perspective for US decision making. *Ecology Letters*, *14*(8), 804–815. https://doi.org/10.1111/j.1461-0248.2011.01631.x

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

De'ath, G., & Fabricius, K. E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, *81*(11), 3178–3192. https://doi.org/10.2307/177409

Ding, S., Chen, M., Gong, M., Fan, X., Qin, B., Xu, H., et al. (2018). Internal phosphorus loading from sediments causes seasonal nitrogen limitation for harmful algal blooms. *Science of the Total Environment*, *625*, 872–884. https://doi.org/10.1016/j.scitotenv.2017.12.348

Fee, E. J. (1979). A relation between lake morphometry and primary productivity and its use in interpreting whole-lake eutrophication experiments. *Limnology & Oceanography*, *24*(3), 401–416. https://doi.org/10.4319/lo.1979.24.3.0401

Frei, R. J., Lawson, G. M., Norris, A. J., Cano, G., Vargas, M. C., Kujanpää, E., et al. (2021). Limited progress in nutrient pollution in the US caused by spatially persistent nutrient sources. *PLoS One*, *16*(11), e0258952. https://doi.org/10.1371/journal.pone.0258952

Geological Survey. (2004). *National Hydrography dataset*. U.S. Dept. of the Interior, U.S. Geological Survey. Retrieved from https://www.usgs.gov/national-hydrography/national-hydrography-dataset

Glibert, P. M., Beusen, A. H., Harrison, J. A., Dürr, H. H., Bouwman, A. F., & Laruelle, G. G. (2018). Changing land-sea-and airscapes: Sources of nutrient pollution affecting habitat suitability for harmful algae. *Global ecology and oceanography of harmful algal blooms*, 53–76. https://doi.org/10.1007/978-3-319-70069-4_4

Glibert, P. M., Pitcher, G. C., Bernard, S., & Li, M. (2018). Advancements and continuing challenges of emerging technologies and tools for detecting harmful algal blooms, their antecedent conditions and toxins, and applications in predictive models. *Global ecology and oceanography of harmful algal blooms*, 339–357. https://doi.org/10.1007/978-3-319-70069-4_18

Gorney, R. M., Graham, J. L., & Murphy, J. C. (2023). The“ H,”“ A,” and“ B” of a HAB: A definitional framework. *Lake Line*, *43*(2), 7–11.

Handler, A. M., Compton, J. E., Hill, R. A., Leibowitz, S. G., & Schaeffer, B. A. (2023). Identifying lakes at risk of toxic cyanobacterial blooms using satellite imagery and field surveys across the United States. *Science of the Total Environment*, *869*, 161784. https://doi.org/10.1016/j.scitotenv.2023.161784

Heisler, J., Glibert, P. M., Burkholder, J. M., Anderson, D. M., Cochlan, W., Dennison, W. C., et al. (2008). Eutrophication and harmful algal blooms: A scientific consensus. *Harmful Algae*, *8*(1), 3–13. https://doi.org/10.1016/j.hal.2008.08.006

Hill, R. A., Fox, E. W., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., & Weber, M. H. (2017). Predictive mapping of the biotic condition of conterminous US rivers and streams. *Ecological Applications*, *27*(8), 2397–2415. https://doi.org/10.1002/eap.1617

Hill, R. A., Weber, M. H., Debbout, R. M., Leibowitz, S. G., & Olsen, A. R. (2018). The Lake-catchment (LakeCat) dataset: Characterizing landscape features for lake basins within the conterminous USA. *Freshwater Science*, *37*(2), 208–221. https://doi.org/10.1086/697966

Ho, J. C., & Michalak, A. M. (2020). Exploring temperature and precipitation impacts on harmful algal blooms across continental U.S. lakes. *Limnology & Oceanography*, *65*(5), 992–1009. https://doi.org/10.1002/lno.11365

Hollister, J. W., & Kreakie, B. J. (2016). Associations between chlorophyll a and various microcystin health advisory concentrations. *F1000Research*, *5*, 151. https://doi.org/10.12688/f1000research.7955.2

Hollister, J. W., Milstead, W. B., & Kreakie, B. J. (2016). Modeling lake trophic state: A random forest approach. *Ecosphere*, *7*(3), e01321. https://doi.org/10.1002/ecs2.1321

Hollister, J. W., Milstead, W. B., & Urrutia, M. A. (2011). Predicting maximum lake depth from surrounding topography. *PLoS One*, *6*(9), e25764. https://doi.org/10.1371/journal.pone.0025764

Iiames, J., Salls, W., Mehaffey, M., Nash, M., Christensen, J., & Schaeffer, B. (2021). Modeling anthropogenic and environmental influences on freshwater harmful algal bloom development detected by MERIS over the central United States. *Water Resources Research*, *57*(10), e2020WR028946. https://doi.org/10.1029/2020wr028946

Isles, P. D., Giles, C. D., Gearhart, T. A., Xu, Y., Druschel, G. K., & Schroth, A. W. (2015). Dynamic internal drivers of a historically severe cyanobacteria bloom in Lake Champlain revealed through comprehensive monitoring. *Journal of Great Lakes Research*, *41*(3), 818–829. https://doi.org/10.1016/j.jglr.2015.06.006

Kicklighter, D. W., Bondeau, A., Schloss, A. L., Kaduk, J., McGuire, A. D., & Intercomparison, T. P. O. T. P. N. M. (1999). Comparing global models of terrestrial net primary productivity (NPP): Global pattern and differentiation by major biomes. *Global Change Biology*, *5*(S1), 16–24. https://doi.org/10.1046/j.1365-2486.1999.00003.x

Kim, J. H., Shin, J.-K., Lee, H., Lee, D. H., Kang, J.-H., Cho, K. H., et al. (2021). Improving the performance of machine learning models for early warning of harmful algal blooms using an adaptive synthetic sampling method. *Water Research*, *207*, 117821. https://doi.org/10.1016/j.watres.2021.117821

Kleinman, P. J., Srinivasan, M., Dell, C. J., Schmidt, J. P., Sharpley, A. N., & Bryant, R. B. (2006). Role of rainfall intensity and hydrology in nutrient transport via surface runoff. *Journal of Environmental Quality*, *35*(4), 1248–1259. https://doi.org/10.2134/jeq2006.0015

Kosten, S., Huszar, V. L., Bécares, E., Costa, L. S., van Donk, E., Hansson, L. A., et al. (2012). Warmer climates boost cyanobacterial dominance in shallow lakes. *Global Change Biology*, *18*(1), 118–126. https://doi.org/10.1111/j.1365-2486.2011.02488.x

Lad, A., Breidenbach, J. D., Su, R. C., Murray, J., Kuang, R., Mascarenhas, A., et al. (2022). As we drink and breathe: Adverse health effects of microcystins and other harmful algal bloom toxins in the liver, gut, lungs and beyond. *Life*, *12*(3), 418. https://doi.org/10.3390/life12030418

Lewis, W. M., Jr. (2011). Global primary production of lakes: 19th Baldi Memorial Lecture. *Inland Waters*, *1*(1), 1–28. https://doi.org/10.5268/iw-1.1.384

Liang, Z., Soranno, P. A., & Wagner, T. (2020). The role of phosphorus and nitrogen on chlorophyll a: Evidence from hundreds of lakes. *Water Research*, *185*, 116236. https://doi.org/10.1016/j.watres.2020.116236

Lin, J., Compton, J. E., Hill, R. A., Herlihy, A. T., Sabo, R. D., Brooks, J. R., et al. (2021). Context is everything: Interacting inputs and landscape characteristics control stream nitrogen. *Environmental Science & Technology*, *55*(12), 7890–7899. https://doi.org/10.1021/acs.est.0c07102

Liu, S., Glamore, W., Tamburic, B., Morrow, A., & Johnson, F. (2022). Remote sensing to detect harmful algal blooms in inland waterbodies. *Science of the Total Environment*, *851*, 158096. https://doi.org/10.1016/j.scitotenv.2022.158096

Loftin, K. A., Graham, J. L., Hilborn, E. D., Lehmann, S. C., Meyer, M. T., Dietze, J. E., & Griffith, C. B. (2016). Cyanotoxins in inland lakes of the United States: Occurrence and potential recreational health risks in the EPA National Lakes Assessment 2007. *Harmful Algae*, *56*, 77–90. https://doi.org/10.1016/j.hal.2016.04.001

Lovett, G. M., & Goodale, C. L. (2011). A new conceptual model of nitrogen saturation based on experimental nitrogen addition to an oak forest. *Ecosystems*, *14*(4), 615–631. https://doi.org/10.1007/s10021-011-9432-z

Marion, J. W., Zhang, F., Cutting, D., & Lee, J. (2017). Associations between county-level land cover classes and cyanobacteria blooms in the United States. *Ecological Engineering*, *108*, 556–563. https://doi.org/10.1016/j.ecoleng.2017.07.032

Meyer, M. F., Topp, S. N., King, T. V., Ladwig, R., Pilla, R. M., Dugan, H. A., et al. (2024). National-scale remotely sensed lake trophic state from 1984 through 2020. *Scientific Data*, *11*(1), 77. https://doi.org/10.1038/s41597-024-02921-0

Milstead, W. B., Hollister, J. W., Moore, R. B., & Walker, H. A. (2013). Estimating summer nutrient concentrations in Northeastern lakes from SPARROW load predictions and modeled lake depth and volume. *PLoS One*, *8*(11), e81457. https://doi.org/10.1371/journal.pone.0081457

Naghdi, K., Moradi, M., Rahimzadegan, M., Kabiri, K., & Tabari, M. R. (2020). Quantitative modeling of cyanobacterial concentration using MODIS imagery in the Southern Caspian Sea. *Journal of Great Lakes Research*, *46*(5), 1251–1261. https://doi.org/10.1016/j.jglr.2020.07.003

NASA Earth Observatory Network Global Maps. (n.d). Retrieved from https://earthobservatory.nasa.gov/global-maps

National Atmospheric Deposition Program. (2020). Total deposition maps. Retrieved from https://nadp.slh.wisc.edu/committees/tdep/

Nietch, C. T., Gains-Germain, L., Lazorchak, J., Keely, S. P., Youngstrom, G., Urichich, E. M., et al. (2022). Development of a risk characterization tool for harmful cyanobacteria blooms on the Ohio river. *Water*, *14*(4), 644. https://doi.org/10.3390/w14040644

Paerl, H. W., Fulton, R. S., Moisander, P. H., & Dyble, J. (2001). Harmful freshwater algal blooms, with an emphasis on cyanobacteria. *The Scientific World Journal*, *1*, 76–113. https://doi.org/10.1100/tsw.2001.16

Pennino, M. J., Leibowitz, S. G., Compton, J. E., Hill, R. A., & Sabo, R. D. (2020). Patterns and predictions of drinking water nitrate violations across the conterminous United States. *Science of the Total Environment*, *722*, 137661. https://doi.org/10.1016/j.scitotenv.2020.137661

Pip, E., & Bowman, L. (2014). Microcystin and algal chlorophyll in relation to nearshore nutrient concentrations in Lake Winnipeg, Canada. *Environment and Pollution*, *3*(2), 36. https://doi.org/10.5539/ep.v3n2p36

PRISM Climate Group, Oregon State University. (n.d.). Retrieved from https://prism.oregonstate.edu

R Development Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Sabo, R. D., Clark, C. M., Bash, J., Sobota, D., Cooter, E., Dobrowolski, J. P., et al. (2019). Decadal shift in nitrogen inputs and fluxes across the contiguous United States: 2002–2012. *Journal of Geophysical Research: Biogeosciences*, *124*(10), 3104–3124. https://doi.org/10.1029/2019jg005110

Sabo, R. D., Clark, C. M., & Compton, J. E. (2021). Considerations when using nutrient inventories to prioritize water quality improvement efforts across the US. *Environmental research communications*, *3*(4), 045005. https://doi.org/10.1088/2515-7620/abf296

Sabo, R. D., Clark, C. M., Gibbs, D. A., Metson, G. S., Todd, M. J., LeDuc, S. D., et al. (2021). Phosphorus inventory for the conterminous United States (2002–2012). *Journal of Geophysical Research: Biogeosciences*, *126*(4), e2020JG005684. https://doi.org/10.1029/2020jg005684

Sabo, R. D., Pickard, B., Lin, J., Washington, B., Clark, C. M., Compton, J. E., et al. (2023). Comparing drivers of spatial variability in US lake and stream phosphorus concentrations. *Journal of Geophysical Research: Biogeosciences*, *128*(8), e2022JG007227. https://doi.org/10.1029/2022jg007227

Sadayappan, K., Kerins, D., Shen, C., & Li, L. (2022). Nitrate concentrations predominantly driven by human, climate, and soil properties in US rivers. *Water Research*, *226*, 119295. https://doi.org/10.1016/j.watres.2022.119295

Scavia, D., Wang, Y.-C., Obenour, D. R., Apostel, A., Basile, S. J., Kalcic, M. M., et al. (2021). Quantifying uncertainty cascading from climate, watershed, and lake models in harmful algal bloom predictions. *Science of the Total Environment*, *759*, 143487. https://doi.org/10.1016/j.scitotenv.2020.143487

Seegers, B. N., Werdell, P. J., Vandermeulen, R. A., Salls, W., Stumpf, R. P., Schaeffer, B. A., et al. (2021). Satellites for long-term monitoring of inland US lakes: The MERIS time series and application for chlorophyll-a. *Remote Sensing of Environment*, *266*, 112685. https://doi.org/10.1016/j.rse.2021.112685

Shen, Q., Li, D., Li, D., Liu, Y., Li, J., & Li, S. (2020). Study on the safe disposal and resource utilization of cyanobacterial bloom biomass in Dianchi Lake, China. *Journal of Applied Phycology*, *32*(2), 1201–1213. https://doi.org/10.1007/s10811-019-01995-3

Shi, K., Zhang, Y., Zhou, Y., Liu, X., Zhu, G., Qin, B., & Gao, G. (2017). Long-term MODIS observations of cyanobacterial dynamics in Lake Taihu: Responses to nutrient enrichment and meteorological factors. *Scientific Reports*, *7*(1), 40326. https://doi.org/10.1038/srep40326

Smith, N. J., Webster, K. E., Rodriguez, L. K., Cheruvelil, K. S., & Soranno, P. A. (2021). *LAGOS-US LOCUS v1.0: Data module of location, identifiers, and physical characteristics of lakes and their watersheds in the conterminous U.S. ver 1*. Environmental Data Initiative.

Søndergaard, M., Larsen, S. E., Jørgensen, T. B., & Jeppesen, E. (2011). Using chlorophyll a and cyanobacteria in the ecological classification of lakes. *Ecological Indicators*, *11*(5), 1403–1412. https://doi.org/10.1016/j.ecolind.2011.03.002

Stachelek, J., Hanly, P. J., & Soranno, P. A. (2022). Imperfect slope measurements drive overestimation in a geometric cone model of lake and reservoir depth. *Inland Waters*, *12*(2), 283–293. https://doi.org/10.1080/20442041.2021.2006553

Stumpf, R. P., Davis, T. W., Wynne, T. T., Graham, J. L., Loftin, K. A., Johengen, T. H., et al. (2016). Challenges for mapping cyanotoxin patterns from remote sensing of cyanobacteria. *Harmful Algae*, *54*, 160–173. https://doi.org/10.1016/j.hal.2016.01.005

Suplee, M. W., Watson, V., Teply, M., & McKee, H. (2009). How green is too green? Public opinion of what constitutes undesirable algae levels in streams 1. *JAWRA Journal of the American Water Resources Association*, *45*(1), 123–140. https://doi.org/10.1111/j.1752-1688.2008.00265.x

Taranu, Z. E., Zurawell, R. W., Pick, F., & Gregory-Eaves, I. (2012). Predicting cyanobacterial dynamics in the face of global change: The importance of scale and environmental context. *Global Change Biology*, *18*(12), 3477–3490. https://doi.org/10.1111/gcb.12015

Tong, Y., Li, J., Qi, M., Zhang, X., Wang, M., Liu, X., et al. (2019). Impacts of water residence time on nitrogen budget of lakes and reservoirs. *Science of the Total Environment*, *646*, 75–83. https://doi.org/10.1016/j.scitotenv.2018.07.255

Topp, S. N., Pavelsky, T. M., Dugan, H. A., Yang, X., Gardner, J., & Ross, M. R. (2021). Shifting patterns of summer lake color phenology in over 26,000 US lakes. *Water Resources Research*, *57*(5), e2020WR029123. https://doi.org/10.1029/2020wr029123

U.S. Department of Agriculture. (2021). *Nass - 2017 Census of agriculture Atlas maps*. USDA National Agricultural Statistics Service. Retrieved from https://www.nass.usda.gov/

U.S. Environmental Protection Agency. (2007). *Survey of the Nation's lakes*. Field Operations Manual. Retrieved from https://www.epa.gov/national-aquatic-resource-surveys

U.S. Environmental Protection Agency. (2010). *National Lakes Assessment 2007 (data and metadata files)*. National Aquatic Resource Surveys. Retrieved from https://www.epa.gov/national-aquatic-resource-surveys

U.S. Environmental Protection Agency. (2011). *2012 National Lakes Assessment*. Field Operations Manual. Retrieved from https://www.epa.gov/national-aquatic-resource-surveys

U.S. Environmental Protection Agency. (2012). *2012 National Lakes Assessment*. Laboratory Operations Manual. Retrieved from https://www.epa.gov/national-aquatic-resource-surveys

U.S. Environmental Protection Agency. (2016a). *National Lakes Assessment 2012: A collaborative survey of lakes in the United States*. U.S. Environmental Protection Agency. Retrieved from https://www.epa.gov/national-aquatic-resource-surveys

U.S. Environmental Protection Agency. (2016b). *National Lakes Assessment 2012 (data and metadata files)*. National Aquatic Resource Surveys. Retrieved from https://www.epa.gov/national-aquatic-resource-surveys

U.S. Environmental Protection Agency. (2023). Harmful algal blooms. Retrieved fromhttps://www.epa.gov/habs

Walsh, E. S., Kreakie, B. J., Cantwell, M. G., & Nacci, D. (2017). A Random Forest approach to predict the spatial distribution of sediment pollution in an estuarine system. *PLoS One*, *12*(7), e0179473. https://doi.org/10.1371/journal.pone.0179473

Wang, L., Liu, L., & Zheng, B. (2013). Eutrophication development and its key regulating factors in a water-supply reservoir in North China. *Journal of Environmental Sciences*, *25*(5), 962–970. https://doi.org/10.1016/s1001-0742(12)60120-x

Wang, R., Goll, D., Balkanski, Y., Hauglustaine, D., Boucher, O., Ciais, P., et al. (2017). Global forest carbon uptake due to nitrogen and phosphorus deposition from 1850 to 2100. *Global Change Biology*, *23*(11), 4854–4872. https://doi.org/10.1111/gcb.13766

Watson, S. B., Miller, C., Arhonditsis, G., Boyer, G. L., Carmichael, W., Charlton, M. N., et al. (2016). The re-eutrophication of Lake Erie: Harmful algal blooms and hypoxia. *Harmful Algae*, *56*, 44–66. https://doi.org/10.1016/j.hal.2016.04.010

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Yuan, L. L., & Jones, J. R. (2020). Rethinking phosphorus–chlorophyll relationships in lakes. *Limnology & Oceanography*, *65*(8), 1847–1857. https://doi.org/10.1002/lno.11422

Yuan, L. L., & Pollard, A. I. (2015). Deriving nutrient targets to prevent excessive cyanobacterial densities in US lakes and reservoirs. *Freshwater Biology*, *60*(9), 1901–1916. https://doi.org/10.1111/fwb.12620

Yuan, L. L., & Pollard, A. I. (2019). Combining national and state data improves predictions of microcystin concentration. *Harmful Algae*, *84*, 75–83. https://doi.org/10.1016/j.hal.2019.02.009

Yuan, L. L., Pollard, A. I., Pather, S., Oliver, J. L., & D'Anglada, L. (2014). Managing microcystin: Identifying national-scale thresholds for total nitrogen and chlorophyll a. *Freshwater Biology*, *59*(9), 1970–1981. https://doi.org/10.1111/fwb.12400

Zhang, J., Cao, P., & Lu, C. (2021). Half-century history of crop nitrogen budget in the conterminous United States: Variations over time, space and crop types. *Global Biogeochemical Cycles*, *35*(10), e2020GB006876. https://doi.org/10.1029/2020gb006876

Zhang, Q. (2018). Synthesis of nutrient and sediment export patterns in the Chesapeake Bay watershed: Complex and non-stationary concentration-discharge relationships. *Science of the Total Environment*, *618*, 1268–1283. https://doi.org/10.1016/j.scitotenv.2017.09.221