Revealing Biotic and Abiotic Controls of Harmful Algal Blooms in a Shallow Subtropical Lake through Statistical Machine Learning

Natalie G. Nelson,[†,‡] Rafael Muñoz-Carpena,*,[†] Edward J. Phlips,[§] David Kaplan,[‖] Peter Sucsy,[⊥] and John Hendrickson[⊥]

[†]Hydrology & Water Quality, Agricultural & Biological Engineering, University of Florida, Gainesville, Florida, United States
[‡]Biological & Agricultural Engineering, North Carolina State University, Raleigh, North Carolina, United States
[§]Fisheries & Aquatic Sciences, School of Forest Resources & Conservation, University of Florida, Gainesville, Florida, United States
[‖]Engineering School of Sustainable Infrastructure and Environment, Environmental Engineering Sciences Department, University of Florida, Gainesville, Florida, United States
[⊥]St. Johns River Water Management District, Palatka, Florida, United States

S Supporting Information

ABSTRACT: Harmful algal blooms are a growing human and environmental health hazard globally. Eco-physiological diversity of the cyanobacteria genera that make up these blooms creates challenges for water managers tasked with controlling the intensity and frequency of blooms, particularly of harmful taxa (e.g., toxin producers, N_2 fixers). Compounding these challenges is the ongoing debate over the efficacy of nutrient management strategies (phosphorus-only versus nitrogen and phosphorus), which increases decision-making uncertainty. To improve our understanding of how different cyanobacteria respond to nutrient levels and other biophysical factors, we analyzed a unique 17 year data set comprising monthly observations of cyanobacteria genera and zooplankton abundances, water quality, and flow in a bloom-impacted, subtropical, flow-through lake in Florida (United States). Using the Random Forests machine learning algorithm, an ensemble modeling approach, we characterized and quantified relationships among environmental conditions and five dominant cyanobacteria genera. Results highlighted nonlinear relationships and critical thresholds between cyanobacteria genera and environmental covariates, the potential for hydrology and temperature to limit the efficacy of cyanobacteria bloom management actions, and the importance of a dual nutrient management strategy for reducing bloom risk in the long term.
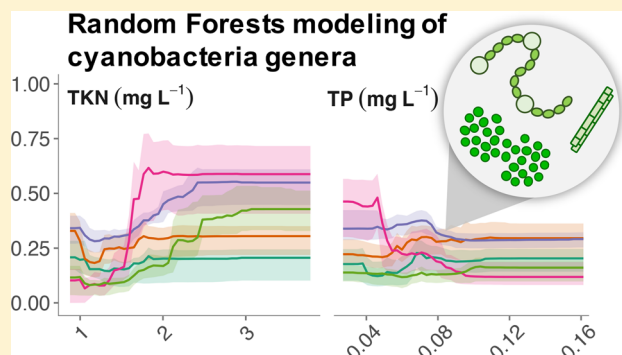
INTRODUCTION

Increases in the frequency, magnitude, duration, and geographic range of harmful cyanobacterial blooms related to cultural eutrophication and climate change[1−3] have raised concerns about undesirable future changes in the structure and function of freshwater ecosystems.[4−8] Of particular concern are cyanobacteria species that threaten ecological and public health through the production of toxins, formation of scums that shade out benthic primary producers, and excess accumulation of biomass that can result in hypoxia.[9]

Although cyanobacteria are found in most aquatic environments, their blooms are particularly widespread and intense in eutrophic freshwater systems.[9] The ubiquity of cyanobacteria is attributable to a diversity of physiological adaptations that allow them to successfully compete for limited resources in a wide range of habitat types.[1,2,10] Such adaptations include N_2 fixation,[11,12] luxury nitrogen (N) and phosphorus (P) uptake,[10,13] buoyancy regulation using gas vesicles,[14] sustained

growth at elevated temperatures,[7] and grazer avoidance,[15,16] among others.

Considering the eco-physiological diversity exhibited by cyanobacteria taxa offers insight into how cyanobacteria outcompete other phytoplankton, but it does not necessarily help clarify the factors that drive the dynamics of cyanobacteria populations and formation of bloom events in individual ecosystems.[17,18] This issue is central to challenges faced by water managers who are tasked with developing appropriate actions for not only controlling the intensity and frequency of blooms but also reducing the potential for blooms of harmful taxa (e.g., toxin producers). These challenges are further compounded by an ongoing debate regarding the efficacy of P-only ("single nutrient") versus N-and-P ("dual nutrient")

management strategies for cyanobacteria bloom abatement in aquatic ecosystems.[19−23]

The primary goal of this study was to address the growing need for increased understanding of the environmental conditions that control blooms of important cyanobacteria genera and to inform water managers' adoption of nutrient management strategies for bloom mitigation. Our specific objectives were to (1) quantify and characterize taxa-specific relationships between cyanobacteria biomass and environmental conditions and (2) evaluate the efficacy of nutrient management strategies relative to several important cyanobacteria taxa. To do so, we applied the Random Forests method, an ensemble-modeling machine learning approach, to a long-term (17 year) data set composed of monthly observations of physical and chemical parameters, flow, and biomass of cyanobacteria and zooplankton taxa in Lake George, a subtropical flow-through lake of the iconic St. Johns River (Florida). To our knowledge, this is the first investigation to concurrently characterize relationships between bloom-forming cyanobacteria genera and biological, chemical, and physical covariates, as well as to apply the Random Forests methodology to the assessment of cyanobacteria dynamics. Results offer new insights into how diverse cyanobacterial assemblages respond to top-down and bottom-up controls over a long-term record, offering insights that are broadly applicable to other freshwater systems.

■ METHODS

Site Description. The St. Johns River is a subtropical blackwater system that drains 24 424 km^2 of Florida's northeastern region (United States).[24] The St. Johns flows from south to north through approximately 500 km of meandering river reaches and flow-through lakes before discharging into the Atlantic Ocean (Figure S1). The largest of these flow-through lakes is Lake George (18 934 ha), a shallow system (mean depth of 3 m) located approximately 210 km upstream from the mouth of the river.[25] Hydrologic inputs to Lake George include the St. Johns River (78.9% of the lake's volume), rainfall (8.7%), flow from artesian springs (8.3%), and runoff (4.1%).[25] Turnover times in the lake are estimated to range from 24 to 180 days depending on flow conditions.[26] Lake George's waters are amber in color because of the presence of high levels of colored dissolved organic matter, concentrations of which peak from fall to early spring.[24]

Lake George is eutrophic (TP concentrations range from approximately 30 to 160 μg L^{-1}) and is affected by recurrent and severe cyanobacteria blooms.[24,27] In terms of carbon biomass, cyanobacteria account for over 90% of the total phytoplankton observed in this system. Cyanobacteria in Lake George display distinct seasonal patterns that diverge from traditional spring−summer−fall−winter variation. Instead, cyanobacteria periodicity in this subtropical, flow-through setting is better explained by "cold" (January−April), "warm" (May−August), and "flushing" (September−December) seasons.[27,28] On average, cyanobacteria carbon biomass in Lake George is greatest in the warm season, followed by the flushing season.[27]

Data Set. Data evaluated in this study included monthly water quality observations, phytoplankton and zooplankton species counts, and flow rates from October 1993 to December 2010. Data were collected at a long-term monitoring site (LG1.2, see Figure S1a) located at the outlet of Lake George. Phytoplankton and zooplankton species identifications and enumerations were done using inverted light microscopes according to methods described by Srifa et al.[27] Water quality parameters associated with the collection of water for plankton counts were obtained from the St. Johns River Water Management District. Regularly measured constituents included alkalinity, chlorophyll, chloride, color, conductivity, dissolved oxygen, pH, pheophytin, Secchi depth, sulfate, total dissolved and suspended solids, total organic carbon, turbidity, water temperature, dissolved ammonium, dissolved phosphate, total phosphorus (TP), and total Kjeldahl nitrogen (TKN). Light extinction due to tripton ($K_{tripton}$, m^{-1}) was predicted from the chlorophyll-a, color, and Secchi depth data (eq S1).

Zooplankton and cyanobacteria biovolumes were transformed into carbon biomass (mg C L^{-1}) using carbon conversion factors of 0.075 and 0.22 pg C μm^{-3}, respectively.[29,30] Daily flow data were collected in the St. Johns River by the U.S. Geological Survey (USGS) at site 02236000 in Deland, FL ("flow monitoring" site in Figure S1a). The USGS maintains a flow monitoring site nearer the inlet to Lake George (site 02236125 in Astor, FL), but these data included several gaps over the study period. Flow data collected at the Deland and Astor sites were compared to ensure that the Deland data were representative of Lake George's inflow dynamics (Figure S2); the data were related by a Pearson's correlation coefficient of 0.96 and a Nash−Sutcliffe efficiency (eq S1) of 0.915, leading us to conclude that the use of Deland data was acceptable for this analysis. Data-processing details are provided in the Supporting Information.

Random Forests Modeling. We utilized Random Forests models to quantify and characterize taxa-specific cyanobacteria−environment relationships. Random Forests is a machine learning algorithm used to fit a large ensemble (or "forest") of randomly assembled decorrelated classification (discrete data) or regression (continuous data) trees to bootstrapped samples of a response variable, and the outputs of these trees are averaged to produce a simulated response.[31−33] The Random Forests modeling framework was selected for this analysis because of its abilities to (1) handle data produced from complex interactions[33] and (2) uncover nonlinear and linear relationship structures,[32,34] making this approach ideal for uncovering the functions relating cyanobacteria genera to environmental and trophic covariates.

While Random Forests models are often viewed as "black boxes," methods exist for quantifying how these statistically derived models relate input explanatory variables to produce simulated responses. Such measures include permutation importance, partial dependence, and relative sensitivity. Permutation importance describes the change in mean squared error (MSE) that occurs when a fitted model is run with a randomly permuted explanatory variable.[34,35] Partial dependence is a measure of an explanatory variable's influence on the response variable given the effects of all other explanatory variables in the model[33,34] and is calculated across the range of each explanatory variable's observations. Relative sensitivity is calculated from the partial dependence curves to summarize the ranges over which partial dependencies varied relative to changes in the explanatory variables; if partial dependence greatly varies over a narrow range of an explanatory variable (i.e., the partial dependence curve has a steep slope), then the response variable is said to have a high relative sensitivity to that explanatory variable.

Models were fit and tested through 5-fold cross validation,[33] which was performed using the following steps: (1) partition
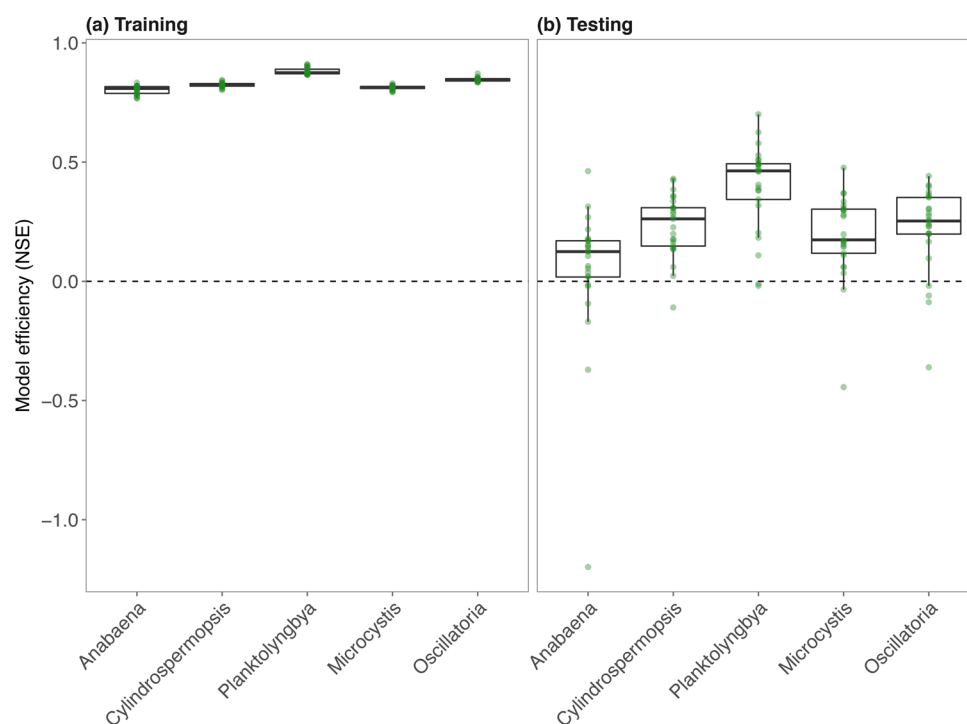
Figure 1. Nash−Sutcliffe Efficiencies (NSE) of the genera models when tested against (a) training and (b) testing data subsets. Each boxplot contains n = 25 values corresponding to the 5-repeated 5-fold cross-validation performed for each Random Forests model; each value is shown as a green point.

data set into training and testing folds; (2) "grow" Random Forests; (3) quantify model performance using the Nash−Sutcliffe efficiency; (4) tabulate permutation importance, partial dependence, and relative sensitivity of all the explanatory variables; and (5) repeat steps 2−4 four times, with each new fold representing the "testing" set on each iteration. This modeling framework was repeated five times for each response variable to evaluate variability in model outputs; thus a total of 25 random forest models were produced with each application (5 repetitions of each 5-fold cross validation). Details associated with each of these steps are provided in the Supporting Information.

Response and Explanatory Variables. The dominant cyanobacteria genera, defined here as genera that accounted for 5% or greater of the total cyanobacteria carbon biomass (mg C L^{-1}) over the study period, were designated as the response variables in the Random Forests modeling analysis.

TKN (mg L^{-1}), TP (mg L^{-1}), dissolved NH$_4$ (mg L^{-1}), dissolved PO$_4$ (mg L^{-1}), water temperature (°C), average flow (m^3 s^{-1}), partial light extinction due to tripton ($K_{tripton}$, m^{-1}), color (CPU), copepods (mg C L^{-1}), and rotifers (mg C L^{-1}) were included as explanatory variables in the Random Forests models. $K_{tripton}$ was calculated from Secchi depth, chlorophyll-a, and color (see the Supporting Information). This reduced set of factors was selected based on reported relationships between these covariates and cyanobacteria.[2,25,27,36−38] Excluded variables are outlined in the Supporting Information, along with justification for their exclusion. All explanatory variables were lagged by one time step (1 month) to reflect the division rates of cyanobacteria in natural environments.[13,39]

■ RESULTS

Summary of Observations. The *Anabaena* (N$_2$ fixers), *Cylindrospermopsis* (N$_2$ fixers), *Planktolyngbya* (nonfixers),

Microcystis (nonfixers), and *Oscillatoria* (nonfixers) groups dominated the Lake George cyanobacteria community over the study period (Table S1). It is important to note that the *Oscillatoria* group includes several genera (e.g., *Planktothrix* and *Limnothrix*) that were further divided from *Oscillatoria* and renamed during the data collection period;[40] similarly, the *Anabaena* group includes members of the genus recently renamed to *Dolichospernum*.[41] Given that these changes occurred well into the study period, we have reported these groups per the prior nomenclature because observations were recorded in the context of these broader classifications.

Cyanobacteria biomass displayed strong interannual (Figure S3) and seasonal variability (Figure S4) across the dominant genera. Of the dominant genera, *Oscillatoria* produced the greatest total biomass over the study period and *Anabaena* the least (Table S1). Interestingly, *Microcystis* regularly bloomed prior to 2001 and minimally varied at low biovolumes thereafter (Figure S3). Cyanobacteria predominantly flourished in the warm season, followed by the flushing season (Figure S4); *Anabaena* typically bloomed prior to the other genera on a seasonal basis (Figure S4). Additional description of the observed data is included in the Supporting Information.

Model Performance. Models fit the training data well, as measured by Nash Sutcliffe efficiency (NSE) across the 25 model applications per response variable (Figure 1a); the minimum NSE value was 0.77 (an *Anabaena* model) and the max was 0.91 (a *Planktolyngbya* model). Previous studies have proposed that models with NSE > 0.65 are "acceptable",[42] indicating that all the training Random Forests models presented here satisfactorily simulated cyanobacteria genera dynamics. However, models did not perform as well when predicting the testing data (Figure 1b). Specifically, the model suites included efficiencies below 0 when evaluated against the testing data, which suggests that findings reported here were
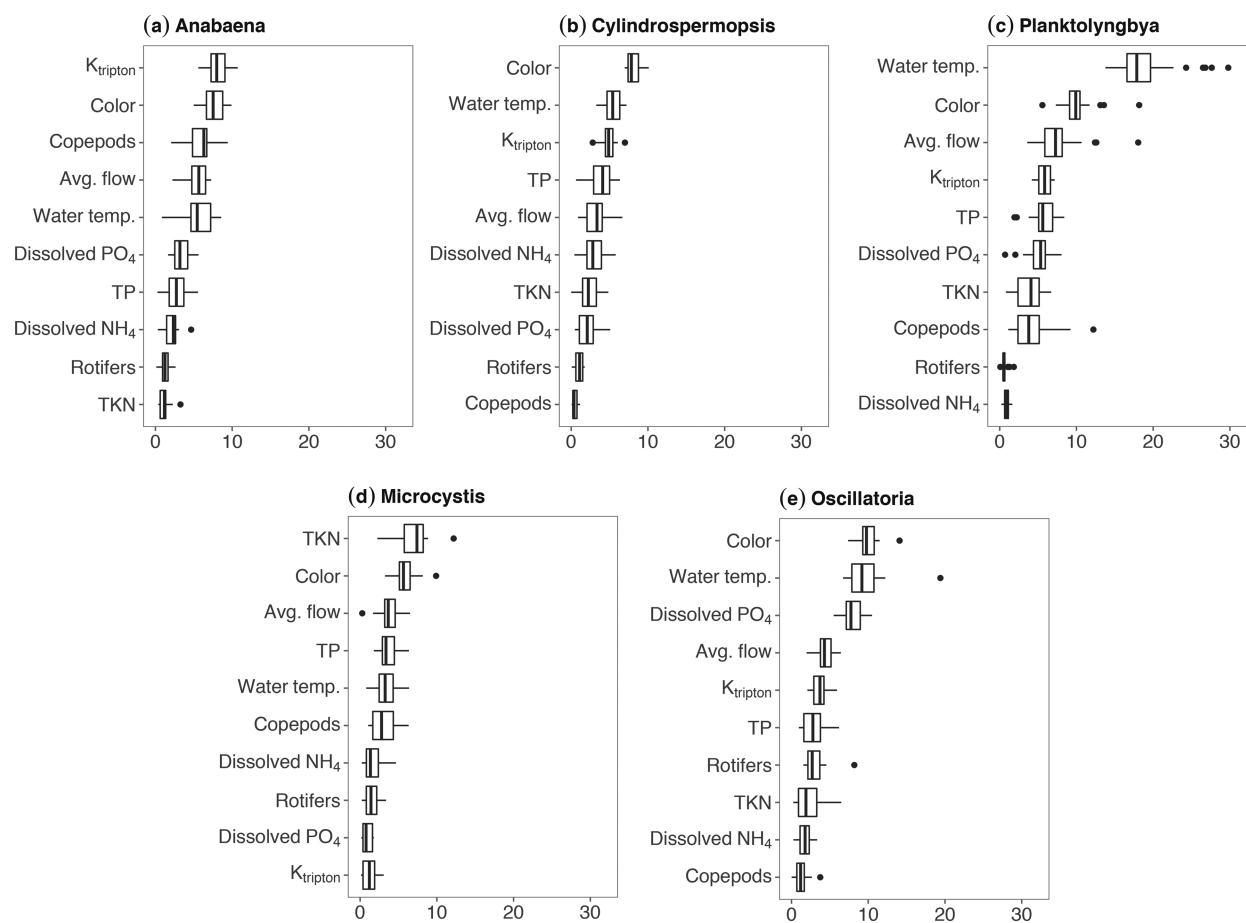
Figure 2. Permuted importance of explanatory variables. *x*-axes are in units of % change in mean squared error. Each boxplot includes *n* = 25 values corresponding to the 5-repeated 5-fold cross-validation performed for each model.

not generalizable across the entire study period; rather, marked effects of particular relationships were important for subsets of the period of record.

Permutation Importance. Color and water temperature were associated with the greatest permutation importance (quantified through percent change in MSE) in the *Cylindrospermopsis* (Figure 2b), *Planktolyngbya* (Figure 2c), and *Oscillatoria* (Figure 2e) models; these explanatory variables were also among the most important in the *Anabaena* (Figure 2a) and *Microcystis* (Figure 2d) models. Overall, permutation of explanatory variables did not dramatically change the percent MSE in any of the models. However, water temperature appeared to be a particularly influential variable in the *Planktolyngbya* model (Figure 2c). As the *Planktolyngbya* model suite had the highest testing NSE results, the variable importance results suggest that permuted importance also served as a predictor of model testing success.
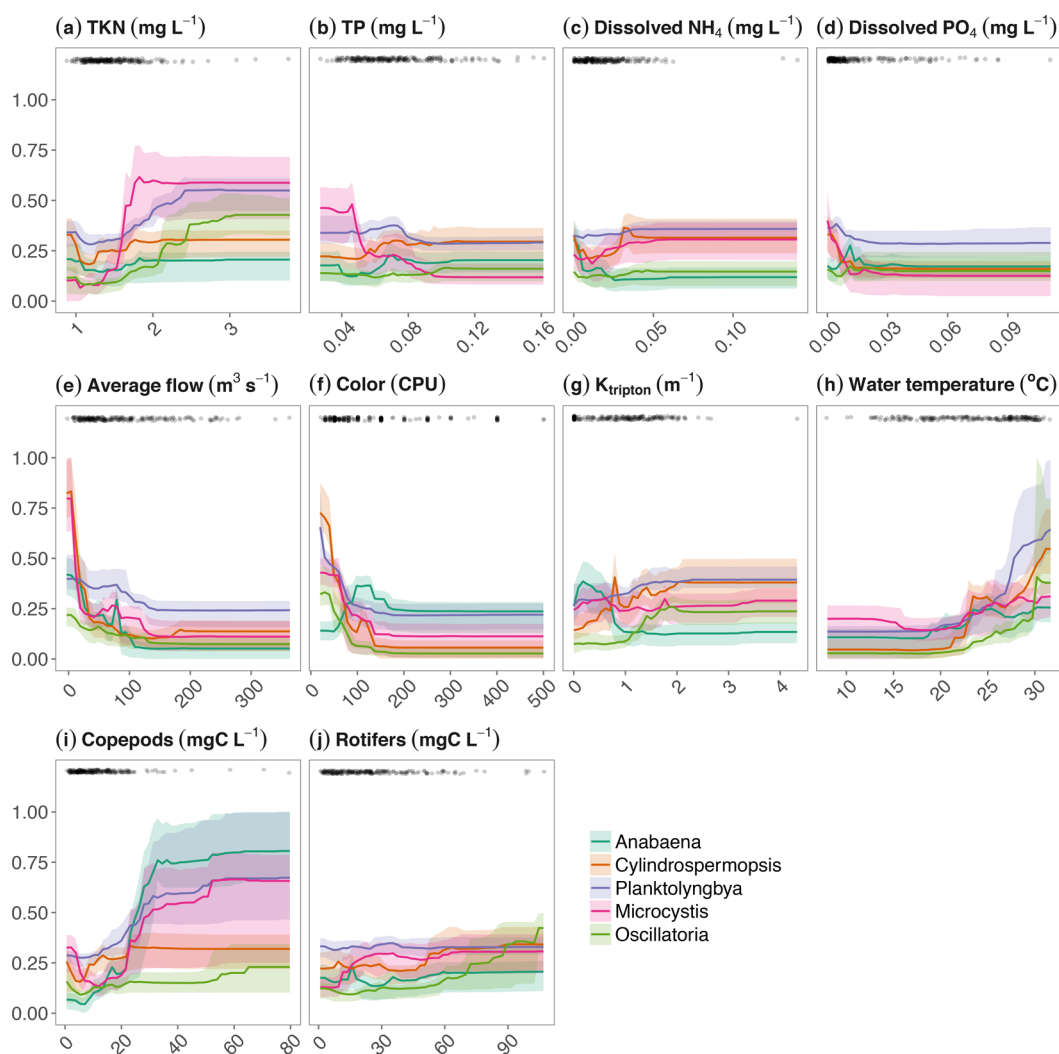
Partial Dependence. Plots of partial dependence revealed predominantly nonlinear model relationships between the lagged explanatory variables and cyanobacteria genera (Figure 3). Partial dependence curves correspond to the relationships linking response and explanatory variables, and the curves are composed of the average modeled values (in units of the response variable) across the range of explanatory variable observations. The steepest curves were associated with TKN, TP, average flow, color, water temperature, and copepod abundance, whereas the curves for NH$_4$, PO$_4$, $K_{tripton}$, and rotifer abundance curves were largely invariant.

Nonfixing cyanobacteria genera (*Planktolyngbya*, *Microcystis*, and *Oscillatoria*) displayed a relatively strong partial dependence on TKN (Figure 3a). Specifically, these curves depict a threshold in the relationship between TKN and nonfixers where partial dependence rose sharply with TKN concentrations between 1.5 and 2.5 mg L^{-1} but plateaued in the range of 2.5–3 mg L^{-1}. Note, however, that there were few observations at high TKN concentrations, limiting interpretation in this range. *Anabaena* and *Cylindrospermopsis* showed little partial dependence on TKN; however, these partial dependence curves revealed a slight sensitivity of N$_2$ fixers to TKN at low TKN concentrations.

The cyanobacteria genera largely lacked partial dependence on TP, with *Microcystis* being an exception (Figure 3b). *Microcystis'* partial dependence on TP indicated a sensitivity to TP at low concentrations (decreasing carbon biomass with increasing TP) up until a concentration of approximately 0.1 mg L^{-1}, above which there was no additional dependence (i.e., similar to the "threshold" response described above).

Average flow (Figure 3e), color (Figure 3f), and water temperature (Figure 3h) emerged as explanatory variables on which all of the cyanobacteria genera displayed notable partial dependence. Specifically, cyanobacteria were more abundant at low flows and colors and higher temperatures. However, the color partial dependence curves were essentially identical across all cyanobacteria genera except *Anabaena* (Figure 3f), which peaked at approximately 120 CPU.

Figure 3. Scaled partial dependence plots of the cyanobacteria genera. Solid lines equal the median partial dependences of the $n = 25$ models per genus, and corresponding shaded areas are bound by the min and max partial dependences. y-axes correspond to carbon biomass, but the curves have been min−max normalized by genus to enable comparisons among the genera, making the y-axes unitless. Partial dependence plots for each genus (nonscaled) are included in the Supporting Information. The point clouds included at the top of each panel reflect the densities of the explanatory covariates' observed values. Units of the x-axes are shown in the sub-plot titles.
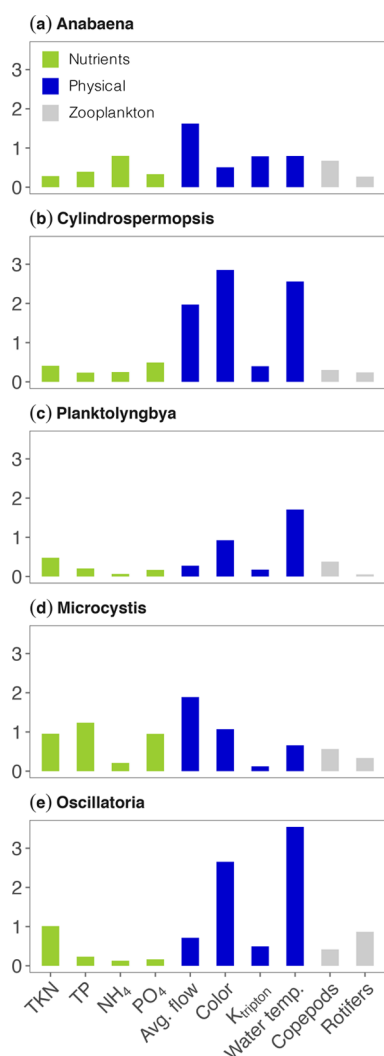
Relative Sensitivity. Relative sensitivities were calculated based on the median scaled partial dependence curves (Figure 4). The cyanobacteria genera were, on average, most sensitive to the physical covariates, followed by nutrient concentrations, then zooplankton abundance (average relative sensitivities to physical factors, nutrients, and zooplankton across all genera equaled 1.29, 0.45, and 0.41, respectively). *Planktolyngbya* (Figure 4c) and *Oscillatoria* (Figure 4e) were most sensitive to changes in water temperature; *Anabaena* (Figure 4a) and *Microcystis* (Figure 4d) were most sensitive to changes in average flow; *Cylindrospermopsis* (Figure 4b) was most sensitive to water color. With regards to nutrients, the genera were generally more sensitive to TKN than TP, though *Microcystis* was similarly sensitive to TKN and TP (Figure 4d).

DISCUSSION

Historically, a number of researchers have argued that nutrient management is the most effective strategy for preventing cyanobacteria blooms regardless of the dominant taxa involved.[17,43,44] While the results of this study provide some support for the potential efficacy of nutrient reduction

strategies, certain aspects of the results highlight how nutrients alone do not always adequately explain trends in biomass of certain cyanobacteria taxa of particular concern from a management perspective (e.g., harmful algal bloom species). The importance of the latter caveat is illustrated by the relative sensitivity of different cyanobacteria genera to the range of explanatory variables included in this modeling effort. Overall, the genera were relatively more sensitive to TKN, TP, average flow, color, and copepod abundance. Because the genera were found to be largely insensitive to changes in NH_4, PO_4, light extinction due to tripton, and rotifer abundance, these covariates were excluded from the ensuing discussion.

Sensitivity of Cyanobacterial Biomass to Nitrogen and Phosphorus Levels. N_2 fixing cyanobacteria were relatively insensitive to TKN (Figure 2a) because of their ability to utilize N_2 as a nitrogen source for growth.[13] In contrast, nonfixing genera showed increasing partial dependence on TKN with increasing concentrations until a threshold of 1.5−2.5 mg L^{-1} (Figures 3a and 2a). This finding indicates that management actions seeking to reduce TKN may produce observable decreases in nonfixer cyanobacteria biomass when

Figure 4. Relative sensitivities of the genera to the explanatory variables. *y*-axes correspond to relative sensitivity values (unitless).

this concentration threshold is crossed. The sharpest thresholding behavior was associated with *Microcystis*, which reflects the observation that after 2001 no major *Microcystis* blooms were observed in Lake George (Figure S3). The post-2001 period was also characterized by lower baseline and peak levels of TKN. It is possible that the relatively strong reliance of *Microcystis* on external nitrogen made it less competitive during the post-2001 period, as observed in some other ecosystems.[45−47]

Most cyanobacteria genera were less sensitive to changes in TP concentrations than to TKN, as demonstrated by the flatter TP partial dependence curves (Figure 3b). The somewhat elevated dependence of *Microcystis* biomass on the lower end of the TP concentration scale may reflect how elevated biomass levels generally appeared in mid-late summer when TP levels in the water column were often lower than in spring or early summer (Figure S5). *Microcystis aeruginosa* is well-known to have strong buoyancy regulation capacity.[13,48] It is possible that during midsummer, when average wind mixing energy is lower than in the spring, *M. aeruginosa* takes advantage of benthic fluxes of bioavailable phosphorus, which are enhanced by lower oxygen levels near the bottom of the water column,[49,50] particularly under high summer water temperatures. The ability of buoyancy-regulating cyanobacteria to take advantage of

elevated nutrient levels in different layers of the water column has been observed in other ecosystems.[51] In more general terms, many of the major bloom-forming cyanobacteria species in Lake George are known to have differing degrees of buoyancy regulation,[52] which may contribute to the apparent low sensitivity to TP levels, because sediment phosphorus reserves can serve as sources of nutrients somewhat independent of water column concentrations.[51]

Implications for Nutrient-Based Cyanobacteria Bloom Mitigation. The cyanobacteria−nutrient relationships identified here provide some insight into the ongoing discussions in the literature of two nutrient management approaches: phosphorus-only[19,20] versus dual control of both nitrogen and phosphorus.[21−23] Although our results are from a single, shallow, subtropical lake, they highlight nuances that are broadly applicable when weighing the potential efficacy of these two approaches. The insensitivity of the cyanobacteria groups' biomass to the range of TP values observed in Lake George raises questions about whether phosphorus reduction alone will significantly reduce intensities of certain types of cyanobacteria blooms. While the results of recent nutrient enrichment bioassay experiments in Lake George identified phosphorus- and phosphorus/nitrogen colimited conditions for phytoplankton growth, these conclusions were produced from data collected during an extreme drought in 2000, when TP concentrations were very low (e.g., near 30 μg L^{-1}).[25] Additionally, the modest sensitivity of N$_2$ fixing species biomass to TP values below 0.06 μg L^{-1} observed in the Random Forests results provides some limited evidence for the potential control of this functional group, possibly related to the relatively high demand some nitrogen fixers have for phosphorus.[53,54] Recent mechanistic modeling efforts by the St. Johns River Water Management District identified an annual mean threshold of 0.05−0.063 mg TP L^{-1} as a long-term target that would be effective for mitigating blooms.[50] The differences between the findings of these prior studies and those presented here are challenging to reconcile but could be explained by the aforementioned influences of internal loads from sediments and water column cycling of nutrients, as observed in some ecosystems.[51] The range of results among this and other studies on cyanobacteria−phosphorus relationships in Lake George warrant further examination, focusing on the feasibility of reducing phosphorus levels within the context of both internal and external phosphorus sources.

In contrast, the sensitivity of non-N$_2$ fixing cyanobacteria to TKN (Figure 3a) suggests that significant reductions in nitrogen load do have the potential to reduce bloom intensities of those genera, though N$_2$-fixing cyanobacteria are unlikely to be affected. Combined with those of TP, these TKN results suggest that the potential exists for control of N$_2$ fixers through phosphorus reductions, and nonfixers through nitrogen reductions. However, other studies have shown that N$_2$ fixation in many lakes may represent less than 50% of the N$_2$ fixer nitrogen demand;[55−57] therefore, these modeling results should be interpreted in the context of other experimental findings.

The value of considering nitrogen load reduction in part depends on the specific management priorities for Lake George. For example, toxic blooms of the non-N$_2$ fixing cyanobacterium *Microcystis aeruginosa* are a recurring problem for Lake George and the downstream portions of the lower St. Johns River.[58] The negative response of *Microcystis* and other non-N$_2$ fixing cyanobacterial biomass to TKN concentrations below 1.5 mg L^{-1} (Figure 3a) suggests that significant

reductions in external nitrogen loads may contribute to less intense blooms of these groups. However, this conclusion may be in part dependent on the response of the N_2 fixing cyanobacteria to reduced nitrogen loads. Currently, the N_2 fixers *Anabaena* (most prominently the renamed genus *Dolichospermum*) and *Cylindrospermopsis* are major contributors to the nitrogen budget of Lake George.[59]

The latter observation raises two important considerations related to possible consequences of nitrogen reduction strategies: (1) The relative importance of *Anabaena* and *Cylindrospermopsis* biomass may increase relative to nonfixers and result in alternative challenges to the ecosystem. (2) Increases in N_2 fixer biomass may lead to increases in nitrogen loads via N_2 fixation to the system, which could ultimately fuel nonfixer blooms, particularly downstream of the lake. These considerations, and those associated with phosphorus reduction scenarios, argue for exploration of a dual phosphorus and nitrogen management strategy. The potential feasibility or efficacy of such an approach in other lakes more generally depends on the specific structure of individual ecosystems and the character of their external and internal sources of nutrient load.

Sensitivity of Cyanobacterial Biomass to Other Abiotic and Biotic Variables. Beyond nutrient-based management strategies, our study also identified important connections among cyanobacteria and other abiotic and biotic factors, which help to put the nutrient response observations into a more holistic management context. One of the challenges facing water managers is that many of these factors (e.g., water temperature, flow, color) are largely uncontrollable in the short term, even though human activities will likely have strong impacts on them in future decades and centuries, such as changes in temperature, rainfall (e.g., as it relates to watershed inputs), and hydrologic conditions (e.g., sea level rise, lake stage). For example, the strong positive relationship between cyanobacteria biomass and temperature also corroborates the hypotheses of a number of researchers that anticipated future global temperature increases will selectively favor cyanobacteria.[4,7,60]

It is also important to examine potential interactive effects among abiotic explanatory variables to further define the extent to which nutrients influence cyanobacteria dynamics, particularly when certain nutrients are not present in limiting concentrations,[17,61−63] as appears to be the case for phosphorus in Lake George. High flow conditions in the fall and winter not only lead to reduced water residence times and reduced light transmission due to elevated color levels but also occur during the period of lowest incident light flux and water temperatures, all of which help to explain the low potential for blooms of phytoplankton in general, despite the fact that it is also the period of highest nutrient concentrations.[24,27,36] The importance of these relationships is reflected in the strong negative correlations between cyanobacterial biomass and both flow and color.

Previous studies of Lake George have highlighted how hydrologic conditions can override or exaggerate the effects of other environmental factors, such as nutrient levels, on cyanobacteria dynamics,[27] corroborating the results presented here. Understanding the degree to which the hydrologic regime decouples cyanobacteria from nutrient dynamics is critical to understanding the potential efficacy of water quality management practices. We caution, however, that there are substantial physical and/or fiscal barriers associated with changing a system's hydrology. Unless managers have the capacity to store and release large amounts of freshwater, their ability to control bloom potentials is likely limited to nutrient management. Therefore, it is ultimately the "nutrient knob"[64] that provides the most reliable management strategy for long-term bloom mitigation.[43,44,64]

Another consideration for defining the dynamics of cyanobacteria populations is the role of top-down controls, both in terms of direct grazing pressure and indirect trophic cascade effects. The data available for the two dominant groups of zooplankton in Lake George, rotifers and copepods, provide some insight into top-down effects. Non-N_2 fixing cyanobacteria had a positive near linear relationship with rotifer biomass, suggesting that rotifers respond positively to increases in cyanobacteria, but do not appear to be able to suppress bloom formation, as indicated by the lack of depression of cyanobacteria biomass even at the highest abundances of rotifers (Figure 3j). The relative resistance of many cyanobacteria to strong top-down control by zooplankton grazers has been widely discussed in the literature, as it relates to the presence of grazing inhibitors,[2,15,16] or secondary trophic effects.[60,65]

For copepods, the relationships to cyanobacteria were positive but nonlinear, with a narrow threshold range of response (Figure 3i). Interpreting the implications of this relationship is complicated by the diverse trophic character of copepods, which include many carnivorous as well as omnivorous taxa.[66] The sharp response threshold for the cyanobacteria−copepod relationship may in part reflect the seasonal coincidence of peaks in copepod and cyanobacteria biomass. In addition, top-down pressure by copepods on smaller zooplankton grazers of phytoplankton (e.g., protozoans, rotifers) may lower grazing pressure on cyanobacteria.

Methodological Implications. Ideally, mechanistic water quality modeling would effectively simulate time-varying dynamics of cyanobacteria; however, the capacity of such models remains limited.[67,68] At present, data-intensive methods (e.g., Random Forests models) are among the most rigorous tools available for the study of complex organism community dynamics using long-term monitoring data. The Random Forests machine learning algorithm utilized here offers a straightforward and assumption-free approach to developing an initial appraisal of how important potentially manageable (e.g., nutrient loads) versus nonmanageable (e.g., water temperature) system features are as predictors of cyanobacteria dynamics.

Although the Random Forests models were employed for explanatory modeling (i.e., describing observed dynamics) in the case presented here, the use of 5-repeated 5-fold cross validation provided insight into the degree to which the models could be used for general prediction purposes. All of the model suites performed reasonably well when applied to testing data (Figure 1b), which suggests that the explanatory variables considered in this analysis are important descriptors of cyanobacteria dynamics. However, some efficiencies associated with individual testing model runs were low (i.e., NSE < 0), which suggests that additional descriptors should be incorporated into the models to increase their predictive capacities. Future research efforts should consider evaluating whether cyanobacteria genera, as opposed to environmental covariates, better predict the dynamics of other genera. Moreover, as this study was limited by the temporal resolution of the long-term monitoring data program, opportunities exist to expand this analysis by repeating it in a setting with higher-resolution data;

given higher-resolution data, the robustness of the sensitivities and nonlinearities reported here could be assessed across a variety of temporal scales.

Applying Random Forests models to a broad variety of aquatic systems impacted by cyanobacteria blooms would help to advance our understanding of the factors that can be managed to mitigate blooms, and such applications should be the focus of future studies. This analysis serves as a case study to guide other researchers and practitioners in their efforts to leverage machine learning tools for the study of water quality phenomena.

■ ASSOCIATED CONTENT

ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.est.7b05884.

> Additional details on data processing, modeling workflow, and summary of observed data (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: carpena@ufl.edu.

ORCID ⓘ

Rafael Muñoz-Carpena: 0000-0003-2838-1514

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

■ REFERENCES

(1) Paerl, H. W.; Otten, T. G. Blooms bite the hand that feeds them. *Science* **2013**, *342* (6157), 433−434.

(2) Paerl, H. W.; Fulton, R. S.; Moisander, P. H.; Dyble, J. Harmful freshwater algal blooms, with an emphasis on cyanobacteria. *Sci. World J.* **2001**, *1*, 76−113.

(3) Glibert, P.; Anderson, D.; Gentien, P.; Granéli, E.; Sellner, K. The Global, Complex Phenomena of Harmful Algal Blooms. *Oceanography* **2005**, *18* (2), 136−147.

(4) Paerl, H. W.; Huisman, J. Climate change: A catalyst for global expansion of harmful cyanobacterial blooms. *Environ. Microbiol. Rep.* **2009**, *1* (1), 27−37.

(5) O'Neil, J. M.; Davis, T. W.; Burford, M. A.; Gobler, C. J. The rise of harmful cyanobacteria blooms: The potential roles of eutrophication and climate change. *Harmful Algae* **2012**, *14*, 313−334.

(6) Carey, C. C.; Ibelings, B. W.; Hoffmann, E. P.; Hamilton, D. P.; Brookes, J. D. Eco-physiological adaptations that favour freshwater cyanobacteria in a changing climate. *Water Res.* **2012**, *46* (5), 1394−1407.

(7) Paerl, H. W.; Huisman, J. Blooms Like It Hot. *Science (Washington, DC, U. S.)* **2008**, *320*, 57−58.

(8) Havens, K.; Paerl, H.; Phlips, E.; Zhu, M.; Beaver, J.; Srifa, A. Extreme Weather Events and Climate Variability Provide a Lens to How Shallow Lakes May Respond to Climate Change. *Water* **2016**, *8* (6), 229.

(9) Paerl, H. W. Nuisance phytoplankton blooms in coastal, estuarine, and inland waters. *Limnol. Oceanogr.* **1988**, *33* (4part2), 823−843.

(10) Oliver, R. L.; Hamilton, D. P.; Brookes, J. D.; Ganf, G. G. Physiology, Blooms and Prediction of Planktonic Cyanobacteria. In *Ecology of Cyanobacteria II: Their Diversity in Space and Time*; Springer: Dordrecht, 2012; pp 1−13.

(11) Howarth, R. W.; Marino, R.; Cole, J. J. Nitrogen fixation in freshwater, estuarine, and marine ecosystems. 1. Rates and importance. *Limnol. Oceanogr.* **1988**, *33*, 688−701.

(12) Howarth, R. W.; Marino, R.; Cole, J. J. Nitrogen fixation in freshwater, estuarine, and marine ecosystems. 2. Biogeochemical control. *Limnol. Oceanogr.* **1988**, *33*, 688−701.

(13) Reynolds, C. *Ecology of Phytoplankton*; Cambridge University Press: Cambridge, U.K., 2006.

(14) Oliver, R. L.; Walsby, A. E. Direct evidence for the role of light-mediated gas vesicle collapse in the buoyancy regulation of Anabaena flos-aquae (cyanobacteria). *Limnol. Oceanogr.* **1984**, *29* (4), 879−886.

(15) Lampert, W. Laboratory studies on zooplankton-cyanobacteria interactions. *N. Z. J. Mar. Freshwater Res.* **1987**, *21*, 483−490.

(16) Wilson, A. E.; Sarnelle, O.; Tillmanns, A. R. Effects of cyanobacterial toxicity and morphology on the population growth of freshwater zooplankton: Meta-analyses of laboratory experiments. *Limnol. Oceanogr.* **2006**, *51* (4), 1915−1924.

(17) Paerl, H. W.; Otten, T. G. Duelling "CyanoHABs": Unravelling the environmental drivers controlling dominance and succession among diazotrophic and non-N2-fixing harmful cyanobacteria. *Environ. Microbiol.* **2016**, *18* (2), 316−324.

(18) Reynolds, C. S.; Huszar, V.; Kruk, C.; Naselli-Flores, L.; Melo, S. Towards a functional classification of the freshwater phytoplankton. *J. Plankton Res.* **2002**, *24* (5), 417−428.

(19) Schindler, D. W.; Hecky, R. E.; Findlay, D. L.; Stainton, M. P.; Parker, B. R.; Paterson, M. J.; Beaty, K. G.; Lyng, M.; Kasian, S. E. M. Eutrophication of lakes cannot be controlled by reducing nitrogen input: Results of a 37-year whole-ecosystem experiment. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (32), 11254−11258.

(20) Schindler, D. W.; Carpenter, S. R.; Chapra, S. C.; Hecky, R. E.; Orihel, D. M. Reducing phosphorus to curb lake eutrophication is a success. *Environ. Sci. Technol.* **2016**, *50* (17), 8923−8929.

(21) Conley, D. J.; Paerl, H. W.; Howarth, R. W.; Boesch, D. F.; Seitzinger, S. P.; Havens, K. E.; Lancelot, C.; Likens, G. E. Controlling eutrophication: nitrogen and phosphorus. *Science (Washington, DC, U. S.)* **2009**, *323* (5917), 1014−1015.

(22) Paerl, H. W. Controlling Eutrophication along the Freshwater−Marine Continuum: Dual Nutrient (N and P) Reductions are Essential. *Estuaries Coasts* **2009**, *32* (4), 593−601.

(23) Paerl, H. W.; Scott, J. T.; McCarthy, M. J.; Newell, S. E.; Gardner, W. S.; Havens, K. E.; Hoffman, D. K.; Wilhelm, S. W.; Wurtsbaugh, W. A. It Takes Two to Tango: When and Where Dual Nutrient (N & P) Reductions Are Needed to Protect Lakes and Downstream Ecosystems. *Environ. Sci. Technol.* **2016**, *50* (20), 10805−10813.

(24) Phlips, E. J.; Cichra, M.; Aldridge, F. J.; Jembeck, J.; Hendrickson, J.; Brody, R. Light availability and variations in phytoplankton standing crops in a nutrient-rich blackwater river. *Limnol. Oceanogr.* **2000**, *45* (4), 916−929.

(25) Piehler, M. F.; Dyble, J.; Moisander, P. H.; Chapman, A. D.; Hendrickson, J.; Paerl, H. W. Interactions between nitrogen dynamics and the phytoplankton community in Lake George, Florida, USA. *Lake Reservoir Manage.* **2009**, *25* (1), 1−14.

(26) Stewart, J.; Sucsy, P.; Hendrickson, J. Meteorological and subsurface factors affecting estuarine conditions within Lake George in the St Johns River, Florida. In *7th International Conference on HydroScience and Engineering*; Philadelphia, PA, 2006.

(27) Srifa, A.; Phlips, E. J.; Cichra, M. F.; Hendrickson, J. C. Phytoplankton dynamics in a subtropical lake dominated by cyanobacteria: cyanobacteria "Like it Hot" and sometimes dry. *Aquat. Ecol.* **2016**, *50*, 163−174.

(28) Srifa, A.; Phlips, E. J.; Hendrickson, J. How many seasons are there in a sub-tropical lake? A multivariate statistical approach to determine seasonality and its application to phytoplankton dynamics. *Limnologica* **2016**, *60*, 39−50.

(29) Latja, R.; Salonen, K. Carbon analysis for the determination of individual biomass of planktonic animals. *Verh. - Int. Ver. Theor. Angew. Limnol.* **1978**, *20*, 2556–2560.

(30) Work, K.; Havens, K.; Sharfstein, B.; East, T. How important is bacterial carbon to planktonic grazers in a turbid, subtropical lake? *J. Plankton Res.* **2005**, *27* (4), 357–372.

(31) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.

(32) Cutler, D. R.; Edwards, T. C.; Beard, K. H.; Cutler, A.; Hess, K. T.; Gibson, J.; Lawler, J. J. Random forests for classification in ecology. *Ecology* **2007**, *88* (11), 2783–2792.

(33) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed; Springer Science & Business Media: New York, 2009.

(34) Jones, Z.; Linder, F. Exploratory Data Analysis using Random Forests. In *73rd Annual MPSA Conference*; 2015; pp 1–31.

(35) Jones, Z. M.; Linder, F. J. Edarf: Exploratory Data Analysis using Random Foresets. *J. Open Source Softw.* **2016**, *1* (6), 92.

(36) Phlips, E. J.; Hendrickson, J.; Quinlan, E. L.; Cichra, M. Meteorological influences on algal bloom potential in a nutrient-rich blackwater river. *Freshwater Biol.* **2007**, *52* (11), 2141–2155.

(37) Leonard, J. A.; Paerl, H. W. Zooplankton community structure, micro-zooplankton grazing impact, and seston energy content in the St. Johns river system, Florida as influenced by the toxic cyanobacterium Cylindrospermopsis raciborskii. *Hydrobiologia* **2005**, *537* (1–3), 89–97.

(38) Paerl, H. W.; Otten, T. G. Harmful Cyanobacterial Blooms: Causes, Consequences, and Controls. *Microb. Ecol.* **2013**, *65*, 995–1010.

(39) Stolte, W.; Garcés, E. Ecological aspects of harmful algal in situ population growth rates. In *Ecology of Harmful Algae*; Springer-Verlag: Berlin, 2006; pp 139–152.

(40) Suda, S.; Watanabe, M. M.; Otsuka, S.; Mahakahant, A.; Yongmanitchai, W.; Nopartnaraporn, N.; Liu, Y.; Day, J. G. Taxonomic revision of water-bloom-forming species of oscillatorioid cyanobacteria. *Int. J. Syst. Evol. Microbiol.* **2002**, *52* (5), 1577–1595.

(41) Wacklin, P.; Hoffmann, L.; Komárek, J. Nomenclatural validation of the genetically revised cyanobacterial genus Dolichospermum (Ralfs ex Bornet et Flahault) comb. nova. *Fottea* **2009**, *9* (1), 59–64.

(42) Ritter, A.; Muñoz-Carpena, R. Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. *J. Hydrol.* **2013**, *480*, 33–45.

(43) Mantzouki, E.; Visser, P. M.; Bormans, M.; Ibelings, B. W. Understanding the key ecological traits of cyanobacteria as a basis for their management and control in changing lakes. *Aquat. Ecol.* **2016**, *50* (3), 333–350.

(44) Rigosi, A.; Carey, C. C.; Ibelings, B. W.; Brookes, J. D. The interaction between climate warming and eutrophication to promote cyanobacteria is dependent on trophic state and varies among taxa. *Limnol. Oceanogr.* **2014**, *59* (1), 99–114.

(45) Gerloff, G. C.; Skoog, F. Nitrogen as a Limiting Factor for the Growth of Microcystis Aeruginosa in Southern Wisconsin Lakes. *Ecology* **1957**, *38* (4), 556–561.

(46) Baldia, S. F.; Evangelista, A. D.; Aralar, E. V.; Santiago, A. E. Nitrogen and phosphorus utilization in the cyanobacterium Microcystis aeruginosa isolated from Laguna de Bay, Philippines. *J. Appl. Phycol.* **2007**, *19*, 607–613.

(47) Peng, G.; Fan, Z.; Wang, X.; Chen, C. Photosynthetic response to nitrogen source and different ratios of nitrogen and phosphorus in toxic cyanobacteria, Microcystis aeruginosa FACHB-905. *J. Limnol.* **2015**, *75* (3), 560–570.

(48) Visser, P. M.; Ibelings, B. W.; Mur, L. R.; Walsby, A. E. The Ecophysiology of the Harmful Cyanobacterium Microcystis. *Harmful Cyanobacteria* **2005**, *3*, 109–142.

(49) Malecki, L. M.; White, J. R.; Reddy, K. R. Nitrogen and Phosphorus Flux Rates from Sediment in the Lower St. Johns River Estuary. *J. Environ. Qual.* **2004**, *33*, 1545–1555.

(50) Hendrickson, J.; Mattson, R.; Sucsy, P. *Recommended Ecosystem Performance Targets to Achieve Designated Use in Lake George, Florida*; Palatka, FL, 2017.

(51) Cottingham, K. L.; Ewing, H. A.; Greer, M. L.; Carey, C. C.; Weathers, K. C. Cyanobacteria as biological drivers of lake nitrogen and phosphorus cycling. *Ecosphere* **2015**, *6* (1), 1–19.

(52) Reynolds, C. S.; Oliver, R. L.; Walsby, A. E. Cyanobacterial dominance: The role of buoyancy regulation in dynamic lake environments. *N. Z. J. Mar. Freshwater Res.* **1987**, *21*, 379–390.

(53) Andersson, A.; Höglander, H.; Karlsson, C.; Huseby, S. Key role of phosphorus and nitrogen in regulating cyanobacterial community composition in the northern Baltic Sea. *Estuarine, Coastal Shelf Sci.* **2015**, *164*, 161–171.

(54) Vahtera, E.; Conley, D. J.; Gustafsson, B. G.; Kuosa, H.; Pitkänen, H.; Savchuk, O. P.; Tamminen, T.; Viitasalo, M.; Voss, M.; Wasmund, N.; Wulff, F. Internal Ecosystem Feedbacks Enhance Nitrogen-fixing Cyanobacteria Blooms and Complicate Management in the Baltic Sea. *Ambio* **2007**, *36* (2–3), 186–194.

(55) Scott, J. T.; Doyle, R. D.; Prochnow, S. J.; White, J. D. Are watershed and lacustrine controls on planktonic N2 fixation hierarchically structured? *Ecol. Appl.* **2008**, *18* (3), 805–819.

(56) Paerl, H. W.; Scott, J. T. Throwing fuel on the fire: Synergistic effects of excessive nitrogen inputs and global warming on harmful algal blooms. *Environ. Sci. Technol.* **2010**, *44* (20), 7756–7758.

(57) Scott, J. T.; McCarthy, M. J. Nitrogen fixation may not balance the nitrogen pool in lakes over timescales relevant to eutrophication management. *Limnol. Oceanogr.* **2010**, *55* (3), 1265–1270.

(58) Hendrickson, J.; Lowe, E. F.; Dobberfuhl, D.; Campbell, D. *Characteristics of Accelerated Eutrophication in the Lower St. Johns River Estuary and Recommended Targets to Achieve Water Quality Goals for the Fulfillment of TMDL and PLRG Objectives*; Palatka, Florirda, 2003.

(59) Doron, M. *Aquatic nitrogen fixation: Patterns, rates and controls in a shallow, subtropical lake*; University of Florida, 2010.

(60) Paerl, H. W.; Paul, V. J. Climate change: Links to global expansion of harmful cyanobacteria. *Water Res.* **2012**, *46* (5), 1349–1363.

(61) Soares, M. C. S.; Rocha, M. I. D. A.; Marinho, M. M.; Azevedo, S. M. F. O.; Branco, C. W. C.; Huszar, V. L. M. Changes in species composition during annual cyanobacterial dominance in a tropical reservoir: physical factors, nutrients and grazing effects. *Aquat. Microb. Ecol.* **2009**, *57*, 137–149.

(62) Arhonditsis, G. B.; Stow, C. a.; Paerl, H. W.; Valdes-Weaver, L. M.; Steinberg, L. J.; Reckhow, K. H. Delineation of the role of nutrient dynamics and hydrologic forcing on phytoplankton patterns along a freshwater-marine continuum. *Ecol. Modell.* **2007**, *208* (2–4), 230–246.

(63) Bormans, M.; Ford, P. W.; Fabbro, L. Spatial and temporal variability in cyanobacterial populations controlled by physical processes. *J. Plankton Res.* **2005**, *27* (1), 61–70.

(64) Paerl, H. W.; Hall, N. S.; Peierls, B. L.; Rossignol, K. L. Evolving Paradigms and Challenges in Estuarine and Coastal Eutrophication Dynamics in a Culturally and Climatically Stressed World. *Estuaries Coasts* **2014**, *37* (2), 243–258.

(65) Sarnelle, O. Herbivore Effects on Phytoplankton Succession in a Eutrophic Lake. *Ecol. Monogr.* **1993**, *63* (2), 129–149.

(66) Kleppel, G. S. On the diets of calanoid copepods. *Mar. Ecol.: Prog. Ser.* **1993**, *99*, 183–195.

(67) Shimoda, Y.; Arhonditsis, G. B. Phytoplankton functional type modelling: Running before we can walk? A critical evaluation of the current state of knowledge. *Ecol. Modell.* **2016**, *320*, 29–43.

(68) Rigosi, A.; Fleenor, W.; Rueda, F. State-of-the-art and recent progress in phytoplankton succession modelling. *Environ. Rev.* **2010**, *18* (NA), 423–440.