Modelling anthropogenic and environmental influences on freshwater harmful algal bloom development detected by MERIS over the Central United States

J. S. Iiames[1], W. Salls[2], M. H. Mehaffey[1], M. S. Nash[1], J. Christensen[2], and Blake Schaeffer[2]
[1] U.S. Environmental Protection Agency, Office of Research and Development, Center for Public Health and Environmental Assessment, USA
[2] U.S. Environmental Protection Agency, Office of Research and Development, Center for Environmental Measurement and Modeling, USA
Corresponding author: John Iiames (iiames.john@epa.gov)

Key Points:

- Cyanobacteria drivers ranked for 369 freshwater lakes in Central United States
- Anthropogenic influences in contributing watersheds showed highest cyanobacteria effects

Abstract

Human and ecological health have been threatened by the increase of cyanobacteria harmful algal blooms (cyanoHABs) in freshwater systems. Successful mitigation of this risk requires understanding the factors driving cyanoHABs at a broad scale. To inform management priorities and decisions, we employed random forest modeling to identify major cyanoHAB drivers in 369 freshwater lakes distributed across 15 upper Midwest states during the 2011 bloom season (July – October). We used Cyanobacteria Index (CI_cyano)—a remotely sensed product derived from the MEdium Resolution Imaging Spectrometer (MERIS) aboard the European Space Agency's Envisat satellite—as the response variable to obtain variable importance metrics for 75 landscape and lake physiographic predictor variables. Lakes were stratified into high and low elevation categories to further focus CI_cyano variable importance identification by anthropogenic and natural influences. 'High elevation' watershed land cover (LC) was primarily forest or natural vegetation, compared with 'low elevation' watersheds LC dominated by anthropogenic landscapes (e.g., agriculture, municipalities, etc.). We used the top ranked 25 RF variables to create a classification and regression tree (CART) for both low and high elevation lake designations to identify variable thresholds for possible management mitigation. Mean CI_cyano was three times larger for 'low elevation' lakes than for 'high elevation' lakes, with both mean values exceeding the 'High' World Health Organization recreational guidance/action level threshold for cyanobacteria (100,000 cells/mL). Agrarian–related variables were prominent across all 369 lakes and low elevation lakes. High elevation lakes showed more influence of lakeside LC than for the low elevation lakes.

Keywords: Cyanobacteria, Satellite, Harmful Algal Blooms, Freshwater, Lakes, Reservoirs, Drivers, VOI, classification and regression tree, CART, random forest, MERIS

*Corresponding author
E-mail address: iiames.john@epa.gov (J.S. Iiames)

1 Introduction

Cyanobacteria harmful algal blooms (CyanoHABs) have proven harmful to human and aquatic health and have devalued recreational (swimming, fishing) and residential uses (Paerl & Otten, 2013; Likens, 1972; Schindler et al., 1975). Algal blooms produced by cyanobacteria can be deemed "harmful" due to the presence of the bloom mass itself, or also in cases where toxins are produced. Untreated algal toxins can cause gastroenteritis, hepatoenteritis and toxic injury to the liver. Carcinoma and other diseases have been linked to several cyanotoxins, including hepatotoxins, neurotoxins, cytotoxins, and dermatoxins (Sivonen 1996; Falconer & Humpage, 1996; Downing et al., 2001). Economic devaluation has resulted from cyanobacterial metabolites producing taste and odor problems that often elude conventional drinking water treatment (Downing et al., 2001). Hypoxic conditions caused by decomposing algal masses have contributed to aquatic degradation in the form of fish kills. In this process, algal masses degrade water clarity which negatively affect invertebrate and fish habitats (Paerl & Otten, 2013). These algal masses have also clogged fish gills in aquaculture systems and have impeded/obstructed water for residential use. In 2007, a bloom in Lake Taihu, China, resulted in a lack of drinking water for nearly one million residents in the nearby city of Wuxi City (Duan et al., 2009). In 2014, a HAB incident in Lake Erie prompted authorities to advise 500,000 people in the Toledo, Ohio area to not use this source water for drinking, bathing, or cooking (Lee, J., 2014). In 2018, drinking water advisories affecting hundreds of thousands of residents were enacted in the city of Salem, Oregon based on elevated levels of cyanotoxins from a bloom on Detroit Lake (The Nowak Consulting Group, 2018). Most recently, a cyanotoxins from a bloom on Clear Lake, affected the drinking water for vulnerable populations (i.e., young children, pregnant women, nursing mothers, people with pre-existing liver conditions, and those receiving dialysis treatment) in the Palm Beach, Florida area (Miller & Sangalang, 2021).

Several environmental conditions (i.e., warming temperatures and water column stratification) in combination with excess nutrient loadings have been linked to cyanoHAB development (Paerl & Huisman, 2008; Michalak et al., 2013). CyanoHABs are not the result of one dominating variable, but the interaction of a multitude of drivers occurring simultaneously (Heisler et al., 2008). These interactions begin on the landscape where nutrients move to the stream networks then eventually the lake. Cyanobacteria requirements for growth are dependent on nutrient and mineral sources, delivery systems, and in-lake processes. Several studies have illustrated conceptual models for cyanobacteria development (Anderson et al., 2015; Knapp & Milewski, 2020).

Both nitrogen (N) and phosphorous (P) have been recognized as key nutrients in the promotion of cyanobacteria formation (Likens, 1972; Schindler, 1977; Paerl, 2008). Nutrient loadings into freshwater systems may originate through fertilization (agricultural and municipal), human and agricultural wastes, soil decomposition and atmospheric deposition (primarily N) (Paerl & Otten, 2013). Even N-rich bedrock may contribute secondarily by providing enhanced growing conditions with eventual biomass decomposition in forested systems (Morford et al., 2011). Agricultural soil decomposition also contributes N loadings, in some instances up to 4.48 kg/ha (Clover, M., 2005). P was seen to be the primary limiting nutrient in a large study of 227 lakes in Canada where N was slowly excluded from these systems - eutrophication continued despite the absence of N (Conley et al., 2009). This study aligns with other earlier research where P was identified as a primary driver in CyanoHAB development (Paerl et al., 2011).

However, these results seem to be predicated on N_2-fixing CyanoHABs within these systems (Paerl et al., 2011).

Nutrient transport to downstream freshwater systems is dependent on several factors which include precipitation, runoff pathways, sediment delivery, soils, flow-route distance and general hydrogeologic scope (Soranno et al., 2015). Sediment and nutrient delivery may be affected by rainfall pattern whether by high intensity-short duration or low intensity-long duration storm events. Rainfall intensity and duration dictates flow path, either by overland flow (high intensity events) or subsurface flow (low intensity events). The period between highly intense rainfall events also can influence the available nutrient run-off into freshwater systems, where longer durations between these events allow for accruing larger loads of N and P from dry atmospheric deposition and mineralization (Kleinman, et al., 2006; Reichwaldt & Ghadouani, 2012; Padilla, 2018). However, the juxtaposition of high rainfall events releasing nutrients into the system to these same events flushing nutrients out of the system may create a net zero sum for the receiving water body (Paerl & Huisman, 2008). Agricultural subsurface tile drainage moves nutrients from fields more effectively than natural drainage accounting for 80% more P and 43% more N than unmanaged systems (Mrdjen et al., 2018; Van Esbroeck et al., 2016; Smith et al., 2015; Drury et al., 1996).

In-lake processes have been studied quite extensively with cyanobacteria correlated with lake-level nutrient concentrations (N and P), light-level (high and low), water temperature (exceeding $20^{o} – 25^{o}$), hydrologic and meteorological conditions, high dissolved organic matter, and sufficient iron (Graham et al., 2016). Hydrologic and meteorological processes determine nutrient delivery, flushing rates and subsequent residence times for nutrient availability (Paerl & Otten, 2013). Calm surface water in conjunction with persistent vertical stratification can create anoxic conditions indirectly determining P nutrient availability affecting release from sediment. Low N and high P (i.e., low N:P ratios) can promote cyanoHAB development (Paerl & Otten, 2013). However, abundance of both N and P does not preclude the development of these cyanoHABs (Paerl & Otten, 2013). Higher water temperatures affect cyanoHAB development through the exploitation of buoyancy regulations (i.e., separation of light and nutrients) in stratified water columns (Paerl & Paul, 2012; Beaver et al., 2014). Seasonal timing of bloom outbreaks may also be linked to increasing global temperatures which increase the strength and depth of the lake stratification, affecting nutrient availability for algal growth (Chirico et al., 2020). Lake morphology is also connected to cyanoHAB development in freshwater systems where mean depth and volume reflect the buffer capacity to nutrient inputs (Taranu et al., 2017; Liu et al., 2011), where smaller lakes with lower water volume were impacted to a greater extent than larger lakes (Mrdjen et al., 2018). Added to the complexity of all these mitigating factors are the observed seasonality effects which govern cyanoHAB development. CyanoHAB development in one system was governed by water temperature and lake discharge (winter), solar radiation and wind (spring), water temperature, solar radiation, and N (summer), and wind (fall) (Mrdjen et al., 2018). Others have commented on seasonality affecting cyanoHAB productivity based on different nutrients becoming available at different points in the season (Boström et al., 1989; Anderson et al., 2002; Elser et al., 2007; Paerl et al., 2011). Finally, in-lake processes and near shore land use and land cover (LULC) tends to influence cyanoHAB development more than in high transport capacity systems, where LULC is influential across the entire watershed (Fraterrigo & Downing, 2008).

To understand the complexity of these interactions, several studies have investigated single lake systems for possible drivers of cyanoHAB occurrences. Typically, only point-in-time

measurements are made over a limited spatial scale, limiting the understanding of bloom drivers to those specific conditions observed in that system. At a regional or continental scale, a lack of *in situ* collected data, both spatially and temporally, challenges our understanding of cyanoHAB propagation. Even where continental scale data is available, temporal limitations may obscure possible linkages to upstream drivers. As an example, the EPA National Lakes Assessment (NLA) dataset is recognized as the most geographically expansive *in situ* lakes dataset developed in the United States (U.S. Environmental Protection Agency, 2010). Here, 1000 lakes greater than 4 hectares in size and greater than 1 meter in depth were sampled in both 2007 and 2012. Yet, these discreet measurements limit analysis to specific point-in-time, species specific assessments of cyanoHAB development. Several studies have utilized this data to connect possible drivers of cyanoHAB development, but under the caveat that the data is limited in the sampling frequency (one to a few samples per season per lake), the timing of the sampling, and the nature of the cyanoHAB species where bloom events occur unevenly in time and space (Wang & Shi, 2008; Duan et al., 2009). Given these sampling limitations and the dynamics existent within the highly variable cyanoHAB species, it is understandable that these driver interactions may be obscured. This was seen in an analysis of the 2012 NLA data where the dependent-independent variables showed very weak trends (Marion et al., 2017), explained by the likelihood that some freshwater systems were sampled during both high and low productivity (Hayes & Vanni, 2018).

Due to the challenges associated with *in situ* data collection, cyanoHAB abundance in freshwater systems has also been assessed from satellite-based platforms with most efforts centered primarily on algorithm development, validation, and improvement (Song et al., 2013; Song et al., 2014; Medina-Cobo et al., 2014; Clark et al., 2017). Spectral albedo, retrieved with a 1–3-day frequency from the European Space Agency's MEdium Resolution Imaging Spectrometer (MERIS) sensor aboard the Envisat satellite instrument, were used to compute the Cyanobacteria Index (CI) from the spectral shape algorithm originally described as optimally oriented between 665 nanometers (nm) and 709 nm (Wynne et al., 2008). The algorithm was further refined to reduce false positives where chlorophytes were confusing the identification of cyanobacteria (Lunetta et al., 2015). This version of the algorithm, termed CI_cyano, uses the CI for biomass estimates, utilizing the spectral shape around the 665 nm band to exclude non-cyan species. Correspondence between CI_cyano and *in situ* cyanobacteria concentration has been reported across the range of cyanoHAB abundance extents spanning 10,000 to > 1 million cells/mL (mean absolute percentage error, MAPE = 28.6%, R^2 = 0.95) (Clark et al., 2017). Strong *in situ* – CI_cyano linkages were observed over a 39-month period (2009 – 2012) with cell counts below 109,000 cells/mL and above 1,000,000 cells/mL (Lunetta et al., 2015). For reference, cell counts above 100,000 cells/mL indicate a World Health Organization (WHO) guidance/action "high" level for cyanobacteria (WHO, 2003). The lack of corroborative *in situ* data between 109,000 mL and 1,000,000 mL resulted in lower correspondence with CI_cyano in this concentration range. In a recent study, CI_cyano was validated by *in situ* presence/absence of the cyanotoxin microcystins and cell abundance data within 30 U.S. lakes in 11 states. Here the algorithm detected cyanoHABs with 84% accuracy (Mishra et al., 2021). This algorithm has also been vetted in Lake Erie (Wynn et al., 2008), for 25 state health advisories nationally (Schaeffer et al., 2018), for expected ecological patterns (Coffer et al., 2020), and national temporal frequency patterns (Coffer et al., 2021).

Satellite-based estimates of CyanoHAB frequency and abundance have enabled research at far greater geographical and temporal scales than provided through *in situ* sampling. Attempts

to identify major drivers in the development of CyanoHAB occurrence within freshwater systems is difficult due to the complexity and interactions at multiple environmental and geomorphological scales. Prior studies have been specific to certain systems – challenging the applicability beyond the ecosystem studied. With this research, we attempt to identify and rank major drivers across a wide geographic range investigating 75 potential drivers which include several nutrient types and sources, delivery systems, and in-lake processes. For the first time, we examine environmental drivers over a large sample size (369 lakes) based on CI_cyano retrievals from the MERIS sensor. We employ a Random Forest (RF) regression tree analysis to identify/rank the most important variables impacting bloom development over the 2011 bloom season (June – September) over 20% of the United States centered about the Great Lakes.

2 Materials and Methods

2.1 Study Area

The study area represents the upper-central region within the United States, an area centered about the Great Lakes and extending from the Dakotas east into the Northeast and south to northern Kentucky. To provide hydrological consistency major drainage areas Vector Processing Units, defined in the National Hydrography Dataset Plus (NHDPlus) Version 2 (McKay et al, 2012), were used to delineate the extent of the study area. The study area extended over eight Vector Processing Units (Northeast, Mid-Atlantic, Great Lakes, Ohio, the Souris-Red-Rainy, and a section of the Upper and Lower Mississippi) and is representative of 20% of the continental United States (CONUS), spanning five of the eight major Omernik Level I Ecoregions. Omernik ecoregions combine both biotic and abiotic occurrences (i.e., geology, wildlife, soils, vegetation, long range meteorology, landforms, land use, and hydrology) that 'reflect differences in ecosystem quality and integrity' (Omernik, 1987; Omernik, 1995). Within this 15-state representation, 70.5% of the sampled lakes resided in three states: Minnesota, Michigan, and Wisconsin. The area was stratified into two areas based on soils, climate, and elevation which followed Omernik Level II designations which we simply labeled 'high' and 'low' elevation (Fig. 1). The 'high elevation' lake/reservoir designations were primarily dominated by forested or natural vegetation whereas the 'low elevation' designations were dominated by anthropogenic landscapes (e.g., agriculture, municipalities, etc.) (Fig. 2). The Omernik 'high' elevation lake designations were: (1) the Atlantic Highlands, (2) the Mixed Wood Shield, and the (3) Ozark, Ouachita-Appalachian Forests. Each of these ecoregions primarily support forest growth over agricultural land use based on terrain, ranging from rugged mountainous (Ozark, Ouachita-Appalachian Forests) to undulating morainal plains and hills, glaciated, irregular plains and wide lacustrine basins, and large sandy outwash plains (Mixed Wood Shield). The 'low' elevation lake/reservoir designation included areas suitable for agricultural development: (1) Southeastern USA Plains, (2) Central USA Plains, (3) Mixed Wood Plains, and (4) Temperate Prairies. Lakes within one Omernik region, the West-central Semiarid Prairie, did not satisfy either designation of anthropogenic or natural landscapes, and were thus included only in the overall ranking assessment.

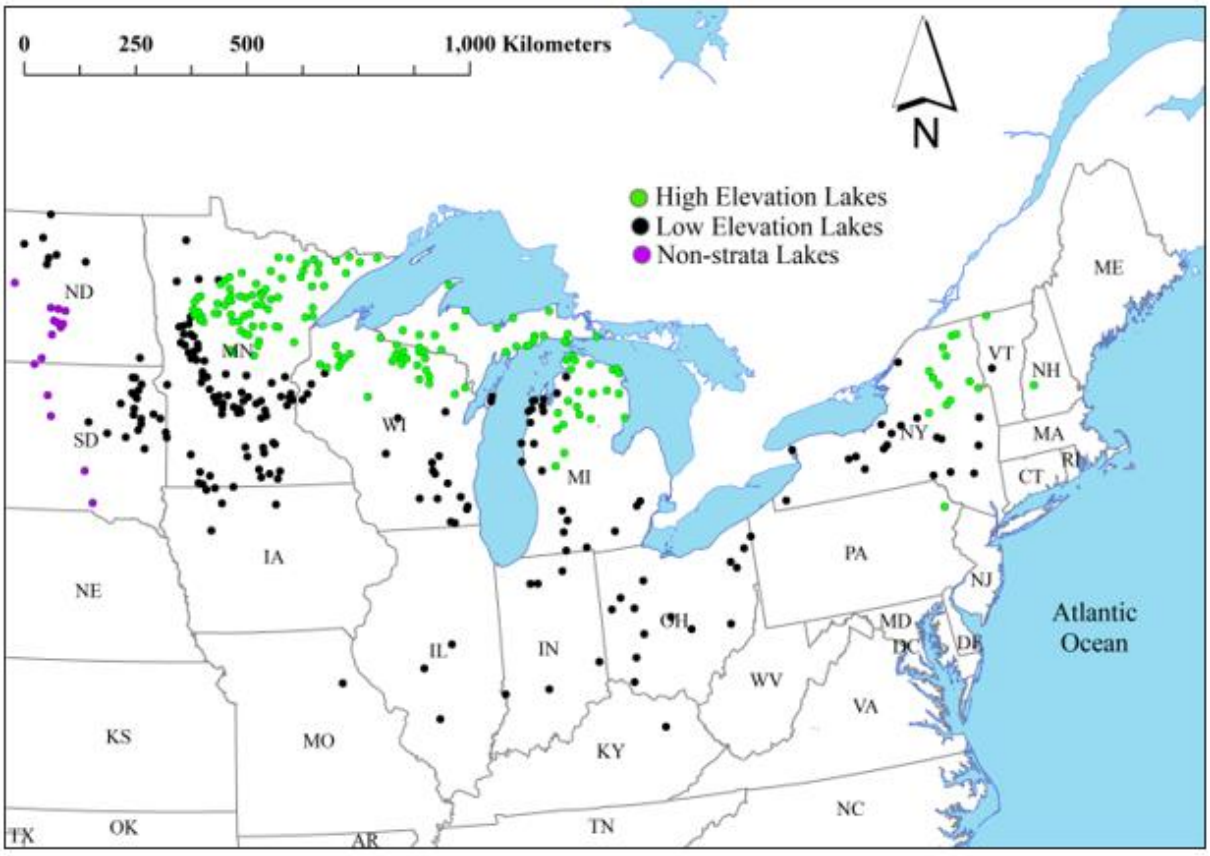


Fig. 1 – Lake distribution showing stratifications based on Omernik ecoregion (high and low elevation)

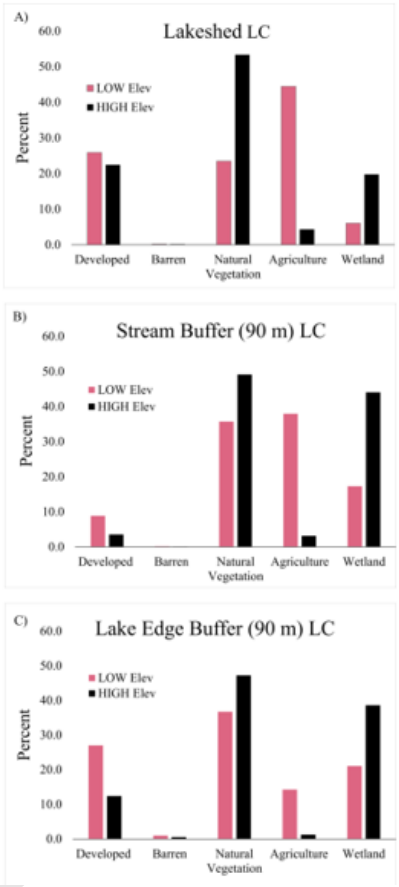


Fig. 2. NLCD 2011 Land cover (LC) percentages across the entire lake watershed (A), within 90 m buffered streams (B), and within 90 m buffered lake edge (C). LC categories are Developed (low, medium, high, open developed), Natural Vegetation (forest [deciduous, coniferous, mixed], herbaceous, and shrub/scrub), Agriculture (row crops, pasture/hay), and Wetland (woody, emergent herbaceous)

2.2. Lake Selection and Lakeshed Delineation

Lakes within our region of interest were selected based on the spatial resolvability by the MERIS sensor. The primary dataset used for lake delineation was the National Hydrography Dataset Plus (NHDPlus) Version 2 (McKay et al., 2012). A lake was deemed resolvable if, at minimum, one 3x3 array of 300 m MERIS pixels (900 m x 900 m) was 100% embedded within the lake boundary, ensuring that at least one MERIS pixel would be 100% water with no spectral land contamination. Once identified, watersheds, or areas contributing to a downslope lake, were delineated around these resolvable lakes. These watersheds herein referenced as ‘lakesheds’ were composed of nested stream catchments, drainage areas contributing to stream segments within a lakeshed. Actual catchment aggregation and hydrological linkages were created through the Lake–Catchment Dataset (LakeCat) (Hill et al., 2018). Lakes with lakesheds extending substantially outside of the study area were removed from the analysis. Nested, overlapping lakesheds also presented an issue; in these cases, land included in multiple lakesheds would be

included in the analysis more than once, providing undue weighting to that area. For each instance of overlap, only the innermost, upstream lake was retained since the link between the landscape in this lakeshed and the lake was thought to be most direct; the relationship is less direct in the case of a large, downstream lake receiving water that first passed through one or more upstream lakes. Filtering all lake and lakeshed criteria yielded 369 lakes stratified into low (n = 187) and high elevation (n = 167) groups for this analysis (note that 15 lakesheds did not meet either the low or high elevation criteria but were included in the non-stratified [i.e., "All"] ranking analysis) (Fig. 1). We investigated these lakes during the bloom season of 2011, which we defined as mid-July through mid-October. We selected the 15-week period spanning July 9 through October 21 to coincide with the timing of satellite data used.

2.3. Ranking and Geospatial Structural Analysis

2.3.1 Random Forest Analysis

RF, an ensemble decision tree model, was implemented to determine variables of importance (VOI) for predictors of cyanoHAB occurrence. The RF algorithm (Breiman, 2001) builds a large collection of decision trees—in this case, regression trees—each constructed by successively splitting data along the variable that minimizes the remaining variance in the response variable at each split. Each tree is based on a random subset of data, which reduces variance and increases predictive power. Additionally, each split within each tree considers only a random subset of the predictor variables, thus decorrelating trees within a random forest and minimizing issues presented by correlation among predictor variables.

Variable importance can be assessed in one of two ways: 1) permutation importance — determined by comparing model error before and after a given predictor's values are randomly rearranged, or 2) mean decrease in impurity— that is, the variance reduction in the response variable from all splits along a given predictor. Permutation importance, selected for our use in this study, essentially amounts to a comparison in model performance between that predictor being present or absent. While this method is somewhat susceptible to correlation among predictors (Strobl et al., 2008), it avoids assigning greater importance to predictors with broader ranges, a known issue with impurity importance (Strobl et al., 2007). Raw importance values have been shown to be preferable to normalized importance values and were therefore used in this study (Strobl and Zeileis, 2008).

Given that high correlation among predictor variables may interfere with proper model fitting and importance assessment, we also investigated the degree of correlation between each pair of variables using a correlation matrix (Moore et al, 2018). RF model is unique in its ability to process correlated data in arriving at ranking solutions (R Core Team, 2013). Across all variables RF can assess the relative value of highly correlated predictors through the process of recursive partitioning where predictors are randomly withheld, including the more powerful predictors. This allows all predictors to be assessed without bias, preventing the enhancement of the more powerful predictors at the expense of the other variables (Stroble et al., 2009; Tomaschek, et al., 2018).

Although RF is capable of handling collinear predictors, strong correlation can interfere with permutation importance, as mentioned. To alleviate this issue, we removed variables sharing correlation coefficient (r) > 0.9 with another variable, in each case retaining the variable

with greater importance based on initial RF importance assessment (Table 1). We fit CI_cyano as the response variable to the suite of 75 predictor variables using the RF package in the R programming language (R Core Team, 2013; Maechler et al., 2013). We used 500 trees considering the recommended value of p/3 predictors at each split, where p is the total number of predictors, resulting in 25. Default settings were used to determine how large, or deep, the trees were grown: minimum size of terminal nodes was five, and no maximum threshold was placed on number of terminal nodes, resulting in relatively deep trees. Since the randomization inherent in the model yields a slightly different forest for each iteration, we generated 100 separate forests and averaged predictor rankings to determine overall ranking. RF contains an embedded means of assessing model performance since each tree considers only a subset of all available observations. The remaining, unused data for each tree are termed out-of-bag (OOB) observations and can be used to independently calculate error for each tree, here in the form of MSE. To assess the ensemble forest, the average MSE across all trees in all forests was calculated. Computational R code for RF analysis is provided.

We also investigated the relationships between the top 10 predictors and CI_cyano for All, High, and Low Lakeshed RF outputs through interpretation of Partial Dependence Plots (PDP). These graphical visualizations in 2-Dimensional space relate the effect on CI_cyano given the marginal effect of a particular predictor (Friedman, 2001).

2.3.2 CART Analysis

In addition, we developed a regression tree using a classification and regression tree (CART) to analyze and predict CI_cyano. CART analysis is a tool that may assist water quality managers to geospatially identify drivers in CI_cyano development with the possibility of mitigating those inputs to reduce impact on these freshwater systems. The tree building technique may reveal hierarchical spatial structure in the relationships between the predictors reflecting the condition of grouped lakes, informing water quality managers with possible mitigation scenarios. RF is limited in providing only the detailed interactions between individual predictors (Tomasckeck, et al., 2018). CART provides the advantage of identifying the most optimal tree given the top VOI inputs from the RF analysis. We ran the CART analysis on the top 25 RF VOIs for the high and low elevation designated lake/reservoir subsets. CART partitions observations into progressively smaller groups to better explain variable interactions, ultimately conditioning on a particular variable – performing a binary split for each independent variable. The independent variable that exhibits maximum homogeneity between the two resulting groups is chosen. Terminal nodes are branch endpoints that imply that further splitting does not yield enough of the variance to be relevant in describing the independent variable of interest (Lawrence & Wright, 2001; Prasad et al., 2006). We ran 100 bootstrap models (trees), allowing the CART algorithm to select the root splitter, then selected the optimal tree (lowest variance) for both high and low elevation designated lakes/reservoirs (Steinberg & Golovnya, 2006).

3. Data Sources

3.1. Response variable: Cyanobacteria Index (CI_cyano)

Full resolution (300 m at nadir) CI_cyano-generated data were acquired from the National Aeronautics and Space Administration (NASA) after processing standard MERIS Level-1B data logged at the NASA Ocean Color website (https://oceandata.sci.gsfc.nasa.gov) (Clark et al., 2017). CI_cyano data is composited weekly, with each pixel value representing the maximum CI_cyano recorded during that week. Pixels with no cloud-free satellite overpasses during a given week are specified as "no data" in the weekly composite. Each weekly CI_cyano composite was masked to the lake boundaries delineated by the NHDPlus data layer. The combination of dynamic *in situ* lake boundaries with possible erroneous NHDPlus lake boundary demarcation, may contaminate lake edge water pixels with spectra from the adjoining land, therefore all lake edge pixels were eliminated for this analysis.

Since each lake required a single CI_cyano value to represent the entire study period, we aggregated CI_cyano values within each lake first across space and subsequently across the time of the study period. Spatially, we considered using maximum CI_cyano for each lake to capture the greatest bloom potential produced by the conditions in and around each lake in each week. However, for the sake of reducing the dominance of individual pixels, we chose the 90th percentile CI_cyano value for spatial aggregation. To aggregate temporally across the 15 weeks of the study period, we calculated median values of the 15 90th percentile CI_cyano values for each lake to represent a typical bloom status during that time. CI_cyano, a unitless index, is reported here in terms of cell abundance through this conversion: Cyano Abundance (cells/mL) = CI_cyano * 10^8 (Lunetta et al., 2015).

3.2. Predictor variables

Predictor variables (n = 75) were selected for this analysis from the literature based on influence on cyanobacteria development or mitigation (Table 1). Sources include LC and physiographic parameters (i.e., slope), lake morphology (depth, volume, area, temperature), nutrient types and delivery (precipitation, anthropogenic impacts), and presence/absence of riparian buffers.

3.2.1. Lake morphology

Lake morphology variables were obtained from a national database generated from the R package lakemorpho (Hollister & Stachelek, 2017). Lakemorpho metrics were calculated using several nationally available datasets such as the National Elevation Dataset and NHDplus. From the available suite of morphological variables, we included mean lake depth, lake volume, and lake surface area as descriptors for each lake.

3.2.2. Precipitation and temperature

Meteorological data were acquired from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) climate dataset (PRISM, 2004). The PRISM dataset provides historical precipitation and temperature data modeled from meteorological stations at a spatial resolution of 4 km across the continental US. PRISM nationwide daily precipitation and

temperature data were downloaded (http://www.prism.oregonstate.edu/explorer/) for the area of interest. Precipitation data was then aggregated into two variables: total seasonal precipitation and maximum precipitation occurring in a 72-hour period. To account for any time lags between a precipitation event and nutrient/sediment entry into the freshwater system, we applied a three-week time lag to our precipitation data, beginning analysis on June 18 three weeks prior to the July 9 date (week 1). To avoid lakeshed size bias and to account for pixels occurring only partially within a lakeshed boundary, we calculated a spatially weighted average for each variable within each lakeshed rather than summing across all pixels.

Since lake water temperature data spanning the entire study area was not available, we calculated mean seasonal air temperature during the study period over each lake as a proxy (Toffolon et al., 2011). As with precipitation, we then calculated a spatially weighted average, doing so across the extent of the lake instead of the entire watershed.

3.2.3. Nutrients

Several edge-of-field nutrient variables were considered to represent application and transport of N and P. Edge-of-field assumes nutrient availability at the field edge for movement out of that field and into an adjacent water body or LC type. The Environmental Policy Integrated Climate (EPIC) model was chosen based on simulation of detailed field-level biogeochemical processes (Williams et al., 2012; Yuan et al., 2018). EPIC simulates an entire suite of N and P process including: (1) N and P losses with sediment, and (2) nitrate N losses in leaching, lateral below surface flow, surface runoff, and tile flow. Inputs include integration of the National Land Cover Database (NLCD) (Yang et al., 2018), county-level crop distribution, and climate variables to produce edge-of-field outputs (kg ha^{-1}) within a 12x12 km grid cell. For each lakeshed, daily values for each variable were summed by pixel to obtain a seasonal total, and then averaged with spatial weighting to obtain one seasonal value to represent the lakeshed.

3.2.4. Land use and physiographic variables

LakeCat provided cumulative upslope catchments summaries for landscape metrics, both anthropogenic and natural landscape features (Hill et al., 2018). Soil composition (percent clay, silt, organic matter), soil erodibility (K-factor), run-off, and depth to water table within LakeCat were summarized from the State Soil Geographic (STATSGO) Database (U.S. Department of Agriculture, 2006). The 2011 NLCD provided the base LC for lakeshed composition.

3.2.5. Stream and Lake buffers

NHDFlowlines were downloaded and extracted from the USGS National Hydrography data portal (https://www.usgs.gov/core-science-systems/ngp/national-hydrography/nhdplus-high-resolution). Streams were clipped to the 369 lakeshed boundaries and 90 m stream buffers were generated using the ArcGIS Riparian Tool within Analytical Tools Interface for Landscape Assessments (ATtILA) (Ebert & Wade, 2004). ATtILA creates the buffers, then calculates NLCD 2011 LC percentages in an iterative process. No internal lake flowlines were buffered.

3.2.6. Artificial Drainage

Spatial maps of agriculture under artificial drainage do not exist apart from state and county estimates (Sugg, 2007). The Soil Survey Geographic database (SSURGO) drainage classes of somewhat poorly drained, poorly drained, and very poorly drained soils specify that poor drainage would restrict or prevent crop production without artificial drainage (Soil Survey Division Staff, 1993). Thus, if LULC indicates that a crop is growing on poorly drained soils, then we assume that the area has likely been modified by artificial drainage to allow for crop production. Areas of poor drainage that intersect with pixels of agricultural crops from the 2012 Cropland Data Layer (CDL) at a 30 m resolution provided a proxy for likely drainage (Christensen et al., 2013; Vanderhoof et al., 2017). The resulting 30 m grid was summarized for each lake catchment, providing the percent of artificial agricultural drainage.

3.2.7. Buffer Distance and Intercepting Sinks

Metrics that link together riparian interception with adjacent agriculture lands were created using a flow-path model originally created for the Chesapeake Bay (Baker et al. 2006) and has been used in subsequent research (Weller et al. 2011, Christensen et al. 2013). The model connects agriculture from the 2006 National Land Cover Database (NLCD) and the 2010 CDL to the existing stream network (Nation Hydrography Dataset – High Resolution 1:24,000-resolution) via flow-paths generated using a 30 m elevation grid (DEM). If the agriculture flow path intersects natural buffers (sinks) in route to the stream, the receiving sinks contiguous with the stream are identified, the width of that buffer distance is counted and resulting buffer widths are assigned to all the original agricultural fields in that flow path. Those buffer widths for each agricultural field were averaged within a lake catchment to calculate the *average distance of agriculture to streams through buffers (m)*. Lake catchments that did not have any agriculture were given the maximum average width value. Likewise, the *percent of sinks that treat agriculture* is the total area of natural area (sinks) that receive flow paths from agriculture lands divided by the total area of natural lands in the lake catchment. Some agricultural areas have flow paths that connect to streams without passing through natural sinks (buffer width of 0 m). The area of non-buffered agriculture within the lake catchment divided by the total area of the lake catchment provides the *percent of agriculture untreated by sinks*. In all these flow-path metrics, agricultural areas include all CDL crop types excluding pasture. Sinks include the natural areas of forest, grassland, and wetland land covers from the 2006 NLCD. Urban classes are ignored in these flow-path calculations.

Table 1.

Cyanobacteria predictor variables.

		Source
Landcover *Overall Lakeshed* *Lake edge 90 m* *Stream buffer 90 m*	*barren, forest (evergreen, mixed);* *developed - open, high intensity;* *pasture; shrub/scrub; grassland/herbaceous;* *open water; wetland* AG on slopes > 10%, >20%;	2011 NLCD
Lake	Mean Lake Depth (m) Lake Surface Area (km^2) Lake Volume (km^3) Ratio of lakeshed to lake area	lakemorpho (NHD+, NED)
	Air temperature (lake temp. surrogate) (C^o)	PRISM (NCAR)
Physiographic	Latitude Avg. dist. of ag. to stream through buffers (m) % of ag. untreated by sink % of sinks that treat ag. % area of ag. occurring on slopes > 10 deg. % area of ag. occurring on slopes > 20 deg. % artificially drained	Derived – 2006 NLCD, 2010 CDL, NHD, DEM NED
Demographic	Housing unit density (housing units km^{-2}) Road-stream intersection density (crossings km^{-2}) Road density (k km^{-2})	USCB
	Density of NPDES sites (sites km^{-2})	NPDES
Meteorological	Precip., max. 72-hour period (cm) Precip., total seasonal (cm)	PRISM (NCAR)
Soil Characteristics	Runoff (mm) Ag soil erodibility (Kfactor index) Soil clay % lithological ferric oxide content % Soil hydraulic conductivity (µm sec^{-1}) Soil erodibility (Kfactor index) Sediment (kg ha^{-1}) Organic matter content % Depth to bedrock (cm) Soil silt % Water table depth (cm) Soil erodibility of ag. land (Kfactor index) Topographic Wetness Index	LakeCat (STATSGO)
Nutrients	Surface N-NH_3, N-NO_3 app. rate (kg ha^{-1}) Sub-surface N-NH_3 app. rate (kg ha^{-1}) lithological N content % N loss in surface runoff (kg ha^{-1}) N in subsurface flow (kg ha^{-1}) Manure application rate (kg ha^{-1}) Inorganic N wet deposition (kg ha^{-1}) Labile P loss in runoff Sediment	EPIC

NED – National Elevation Dataset; CDL – Crop Data Layer (USDA); NHD – National Hydrography Dataset; USCB – US Census Bureau (Tiger/Line - Roads); NPDES – National Pollutant Discharge Elimination System; NLCD – National Land Cover Dataset; PRISM – Parameter-elevation Regressions on Independent Slopes Model; EPIC – Environmental Policy Integrated Climate; STASGO – State Soil Geographic Database; NCAR – National Center for Atmospheric Research

4 Results and Discussion

4.1. CI_cyano and Variable Descriptive Statistics and Correlations

Low elevation lake mean CI_cyano cell counts were approximately three times larger than that of the high elevation lakes (Table 2; Tables S1, S2). Also, 59.4% of low elevation lakes exceeded the WHO guidance/action "high" level (>100,000 cells/mL) compared to 41.3% of the high elevation lakes (WHO, 2003). The largest difference between the two strata was seen with lakes exceeding 1 million cells/mL (20.9% – low, 0.6% – high).

Thirty-seven of the original 75 input variables were LC proportions derived from the 2011 NLCD distributed within a 90 m lake edge buffer (n = 12), 90 m riparian buffer (n = 12), and total lakeshed area (n = 9). The agriculture and developed land proportion in the low elevation lakes/reservoirs exceeded that within the high elevation lakes/reservoirs, 70.3% to 26.7% respectively. This disparity held for both the riparian (low elevation – 46.8%; high elevation – 6.7%) and the lake edge (low elevation – 41.3%; high elevation – 13.6%) buffers (Fig. 2).

With the original 75 predictor variables, there were 2,775 unique pairwise combinations. A correlation matrix for all 75 variables across all 369 lakesheds resulted in 1.7% of the pairings labeled as 'highly' correlated, i.e. $r > 0.70$ (Moore, et al, 2018). The majority (83.2%) of the correlations fell below $r = 0.30$, indicating very weak relationships. Stratification of lakesheds into high and low elevation did not change the percentages substantially (Table 3). Highly correlated pairings were expected based on expert knowledge of the relationships and variable inputs into other derived variables (i.e., 2011 NLCD crop percent as input to nutrient edge of field variables). These pairings included lake physiographic conditions (i.e., lake surface area to lake volume), nutrient loadings (nutrient application rates to percent row crops), agricultural soil conditions (agricultural soil erosion to percent row crop, percent clay, and edge of field N and P), and artificial drainage (agricultural tiling to nitrogen application rates).

Table 2.
CI_cyano cell abundance (c/mL = cells/mL) descriptive statistics for "All", "high elevation", and "low elevation" lakes.

Stratum	n	Mean (c/mL)	Median (c/mL)	S.D. (c/mL)	Min (c/mL)	Max (c/mL)	%Lakes 100k-500k (c/mL)	%Lakes 500k-1 mil. (c/mL)	%Lakes >1 mil. (c/mL)
All	369	367,615	110,267	565,531	10,000	3,019,952	28.5	11.9	11.9
High	167	140,406	76,239	185,330	10,000	1,023,293	34.1	6.6	0.6
Low	187	535,931	251,183	682,001	10,000	3,019,952	23.5	15.0	20.9

Table 3.
Percent of lakes (All, high elevation, low elevation) and number of unique pairwise combinations with predictor variables for three binned ranges.

R	All	High Elevation	Low Elevation
>0.70	46 (1.7%)	31 (1.1%)	30 (1.1%)
0.30 – 0.70	419 (15.1%)	362 (13.0%)	385 (13.9%)
<0.30	2307 (83.2%)	2382 (85.8%)	2360 (85.0%)

4.2. Random Forest Model – 25 VOI

4.2.1. RF All lakes/reservoirs

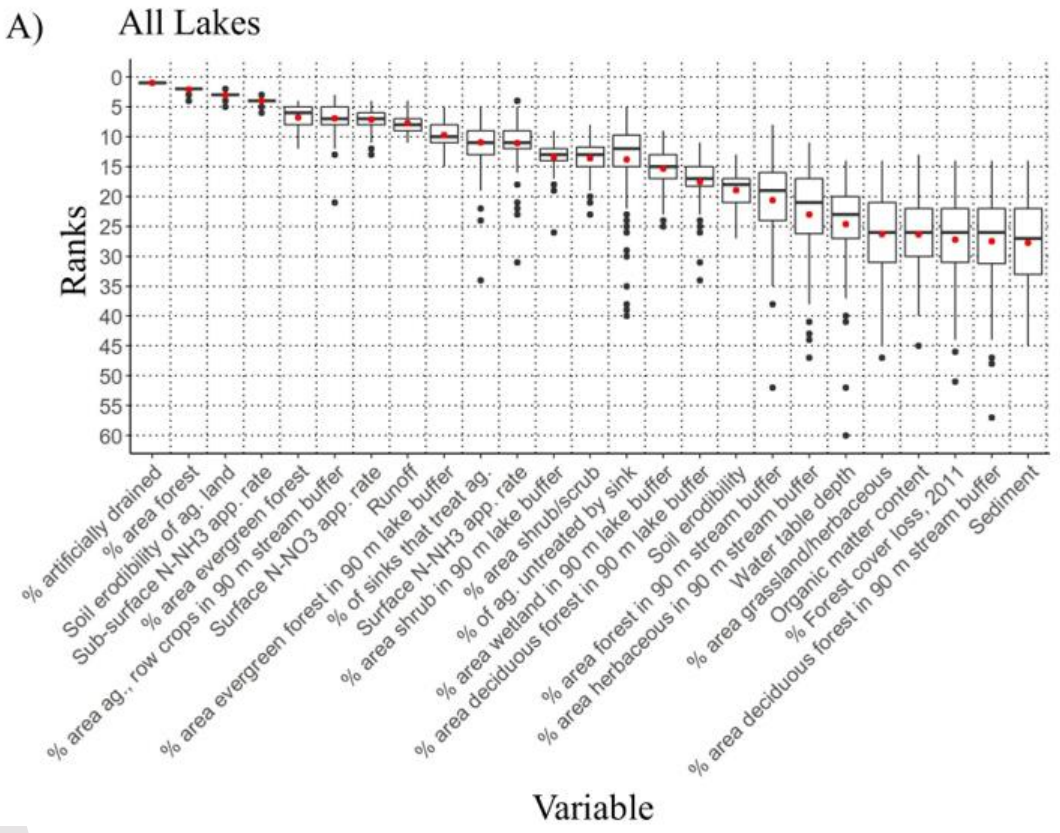
Across the 100 model runs, the variability in predictor rank tended to be higher among predictors with lower ranks (i.e., less important to the model; Fig. 3A). This indicates that the more highly ranked variables were appearing in the top-most ranking throughout most of the 100 model runs. The top ten predictors were primarily agricultural-based (6 of 10) (Table 4) with PDP's indicating two of the top ten predictors had a negative effect on CI_cyano development. Despite only 31.7% of the 369 lakesheds showing a presence of artificial (i.e., 'tile') drainage, this top ranked predictor, showed a sharp spike in CI_cyano development from no artificial drain presence to some presence of this predictor within a lakeshed (Fig. 4A). This corresponds to studies of tile drainage effects on small and large freshwater systems, where the receiving water bodies showed elevated concentrations of nitrate and soluble reactive phosphate and total phosphorous loads, precursors to cyanoHAB development (Mrdjen et al., 2018; Michaud et al., 2019). Agricultural soil erodibility, positively correlated with CI_cyano development, did not show a CI_cyano increase response until the K-factor exceeded 0.15 (Fig. 4B). This may be explained where some cyanobacteria have demonstrated the ability to colonize unstable sandy soils and self-assemble into twisted filaments creating matted-like structures (Garcia-Pichel & Wojciechowski, 2009). Runoff showed an initial negative impact on CI_cyano development before exhibiting a slight positive correlation after approximately 200 mm (Fig. 4C). This may be explained initially with the flushing of stagnant water out of the lake and the mixing of lake thermal layers – both mechanisms offsetting the effect of nutrient transport from the landscape into the lake (Reichwald & Ghadouani, 2012). However, longer lag times may produce blooms later in the year beyond our seasonal time frame. Average CI_cyano MSE across model runs was 2.63e-05 (Table 5). To give a more direct sense of the magnitude of this error relative to CI_cyano response values, the RMSE was 5.12e-03. This relatively high error renders the model unfit for prediction; however, this degree of error did not preclude identification of the most important variables, as supported by the low variability in rank of highly important variables across the ensemble forest. Table 3 lists the top 10 ranks for each lake category and the associated ranks of the other lake categories.

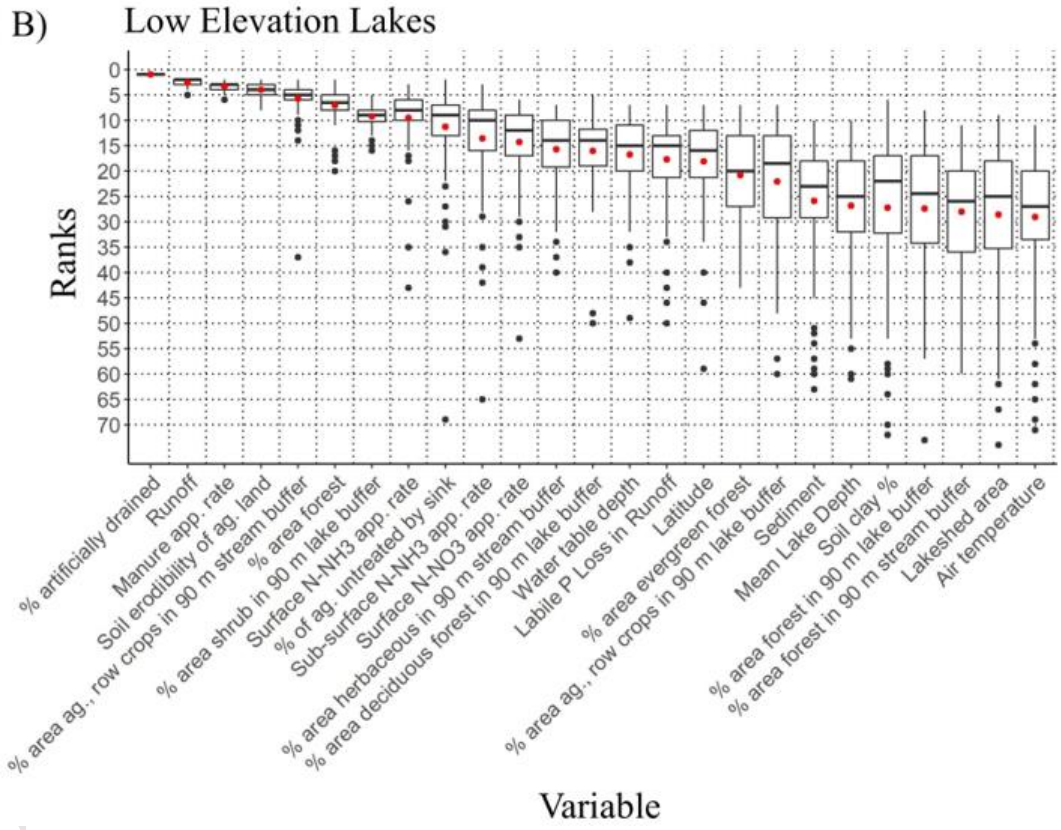
Table 4.
Top 10 predictor variables by stratification inhibiting (r < 0) or advancing (r > 0) CI_cyano development in freshwater systems.

	Predictors	r	All	Low	High
Landcover	% area evergr. forest	-0.25	5		
	% area forest	-0.40	2	6	
	% area decid. forest in 90m lake buffer	-0.16			10
	% area evergr. forest in 90m lake buffer	0.31	9		
	% area shrub in 90m lake buffer	-0.02		7	
	% area wetland in 90m lake buffer	-0.08			7
	% area row crop in 90m strm buffer	0.42	6	5	
	% area developed (low-med) in 90m lake buff	0.04			8
	% area wetland	-0.18			2
Lake	Lake depth (mean)	-0.22			6
Physiographic	% artificial drainage - agriculture	0.11	1	1	
	% of ag. untreated by sink	0.43		9	
	% of sinks that treat ag.	0.42	10		
	Runoff	-0.25	8	2	
Demographic	Road density	-0.03			9
	Housing Unit Density	-0.06			5
Soil	Soil erodibility - Ag	0.47	3	4	
	Water table depth	-0.09			3
	Soil erodibility	0.27			4
	Organic matter content	-0.12			1
Nutrients	Sub-surface N-NH_3 app. rate	0.46	4	10	
	Surface N-NH_3 app. rate	0.46		8	
	Manure app. rate	0.32		3	
	Surface N-NO_3 app. rate	0.42	7		

Table 5.
Summary statistics of out-of-bag mean square error (MSE) across all 100 RF runs for each of the study area sets. Note: RMSE and MSE = CI_cyano.

Statistic	All	Low Elevation	High Elevation
MSE	2.63e-05	4.37e-05	3.51e-06
RMSE	5.12e-03	6.61e-03	1.87e-03
Mean (cells/mL)	3.68e+05	5.36e+05	1.40e+05
Median (cells/mL)	1.10e+05	2.51e+05	7.62e+05
Min (cells/mL)	1.00e+04	1.00e+04	1.00e+04
Max (cells/mL)	3.25e+06	3.02e+06	1.02e+06
SD (cells/mL)	5.66e+05	6.82e+05	1.85e+05





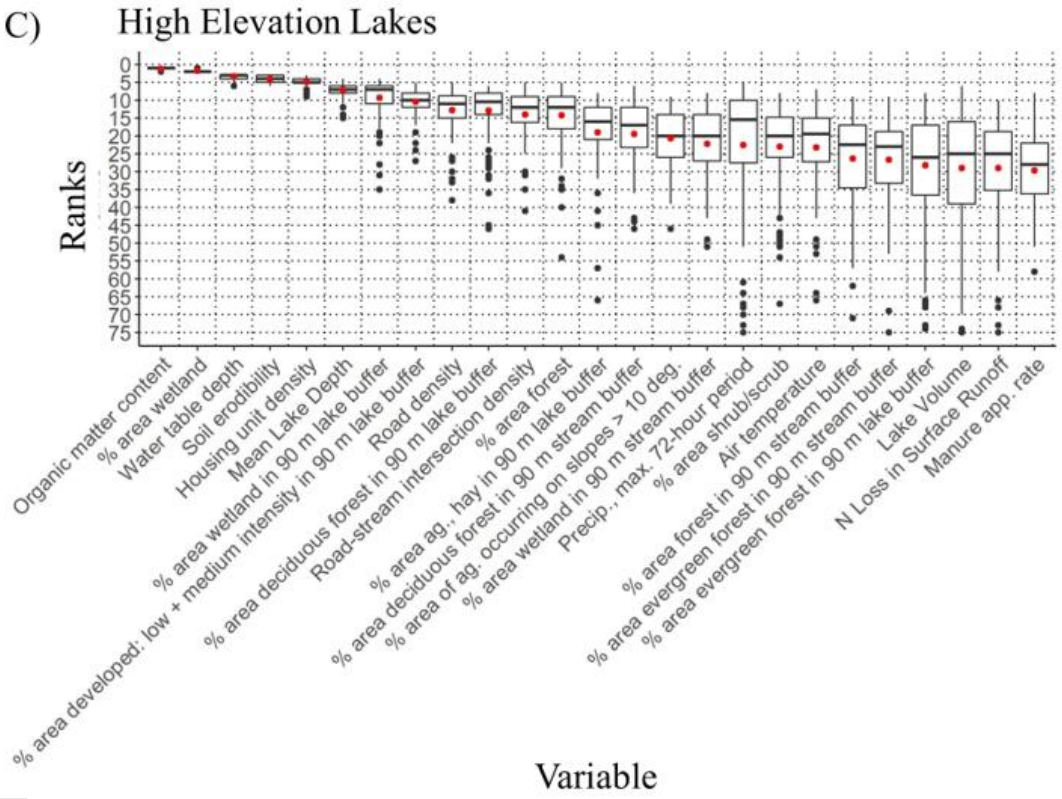


Fig. 3. Distribution of predictor VOI ranks across RF model runs (*n* = 100) for (A) all lakes (n = 369), the (B) low (n = 187) and the (C) high (n = 167) elevation lakes. Variables in each plot are sorted by mean rank, which is represented by red dots for each variable. Box center bar shows median rank, box extents show the first and third quartiles, and whiskers extend to the farthest value within 1.5 x IQR from the first and third quartile respectively; values outside this range are shown as black dots.

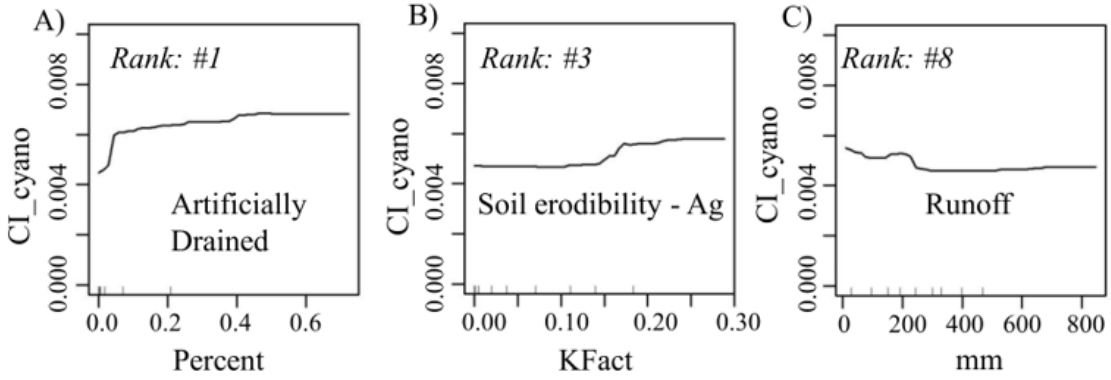


Fig. 4. Examples of partial distribution plots (PDPs) generated from the 100 random forest runs for all 369 lakesheds.

4.2.2. RF Low Elevation lakes/reservoirs designation

RF model runs (*n* = 100) for the 187 lakes/reservoirs in low elevation designations yielded rankings similar to those for the entire study area (i.e., all 369 lakes), with seven agricultural variables ranked in the top ten (Fig. 3B, Table 4). Artificially drained area was again the most important variable, followed by runoff, agricultural soil erodibility, manure application rate, and row crops in a 90 m stream buffer. This trend of agricultural associations with cyanoHAB development has been observed in areas artificially drained (Mrdjen et al., 2018). Also, nearly 2/3's of the NLA lakes/reservoirs sampled in 2007 with high microcystin concentrations were in crop dominated areas, also corresponding to our study results (Beaver et al., 2014). PDPs of these agricultural inputs also showed the largest spikes in CI_cyano development when compared to other inputs (Figs. 5A and 5C). Shrub buffers adjacent to the lake edge showed a negative impact on water quality (Fig. 5B). In comparison, treed buffers provide rooting gaps for water collection and when coupled with a thick understory of shrubs and grasses also provide obstruction to effectively slow water movement over the landscape (Bentrup, 2008). Here, we posit that a purely shrub buffer possibly lacked the rooting zone provided within forested systems and the obstruction capabilities provided by a grass layer (Fig. 5B). MSE across model runs was slightly higher than for the entire study area at 4.37e-05 with RMSE at 6.61e-03 (Table 4).

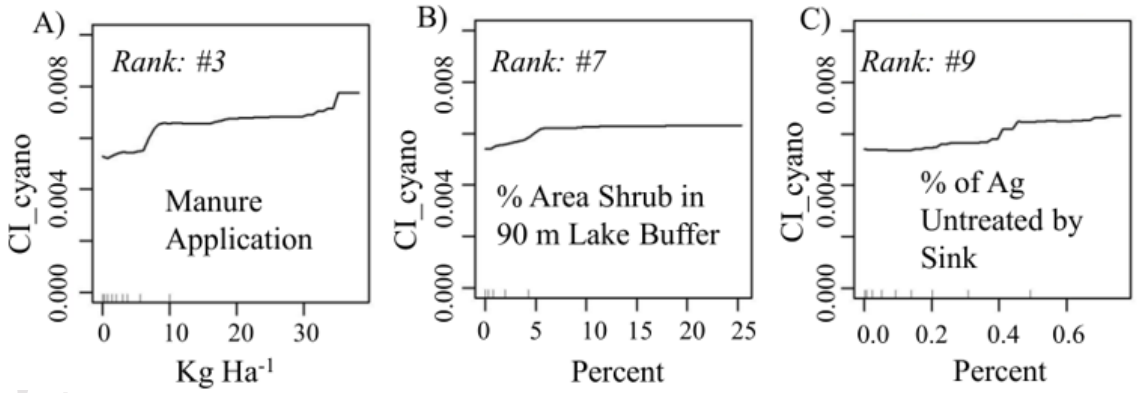


Fig. 5. Examples of partial distribution plots (PDPs) generated from the 100 random forest runs for "Low Elevation" (n = 187) lakesheds.

4.2.3. RF High Elevation lakes/reservoirs designation

RF model runs (*n* = 100) for the 167 lakes/reservoirs in the high elevation designations showed markedly different patterns from low level lakes. A comparison of the variable rankings for the high elevation lakes/reservoirs and the lakes within the entire study area shared 10 of the top 25 variables (none in the top 10) (Fig. 3C, Table 3). Organic matter ranked highest within these lakes, followed by percent wetland area within the lakeshed, water table depth, soil erodibility, and housing unit density. Correlations between CI_cyano and these top variables were generally weaker than for the entire study area and for the low elevation designation and were primarily negatively correlated (8 of 10) (Table 4). Negative correlations with road and

housing unit density made sense if both replaced land out of other positively correlated LC types (i.e., agricultural land) (Figs. 6A and Fig. 6B). However, we would expect that the increase in urbanization with associated increases in impervious surface area and storm drainage systems would contribute to nutrient increases within freshwater systems. Along with decreased infiltration, nutrients can bypass biologically active zones that normally would act to retain or remove these nutrients from the system (Bell et al., 2019). The increasing presence of deciduous forest within a 90 m lake buffer tended to decrease the presence of CI_cyano (Fig. 6C). However, a slight uptick in CI_cyano presence can be seen when deciduous forest increased beyond 40%. This effect could be the result of a light-limiting upper-story canopy, conditions possibly inhibiting the presence of understory species (i.e., grasses and shrubs) necessary to slow overland water flow. Strong storm event precipitation (VOI rank = 17) was associated with lower CI_cyano, a possible indication of flushing the freshwater system of acquired nutrients and mixing of thermal layers. Water table depth (VOI rank = 3) was negatively correlated with CI_cyano development possibly due to the presence of a larger aeration zone available for storm water infiltration (Remson et al., 1959). Also, deeper water tables may follow drier conditions based on precipitation fluctuations between seasons and years. Soil erodibility (VOI rank = 4) in conjunction with runoff can increase the movement of P into a freshwater system. Through the process of adsorption, P available in soil solution can attach to soil particles, binding primarily on clay surfaces and iron and aluminum oxides and can therefore be transported to these waters, providing this nutrient available for cyan development following desorption (Weihrauch & Opp, 2018). The soils found in this high elevation lakeshed stratum had ranks of silt (VOI = 29) and clay (VOI = 31), mineral surfaces preferable for P adsorption. MSE across model runs was 3.51e-06, slightly lower than for the entire study area and low elevation designation; RMSE was 1.87e-03 (Table 4).

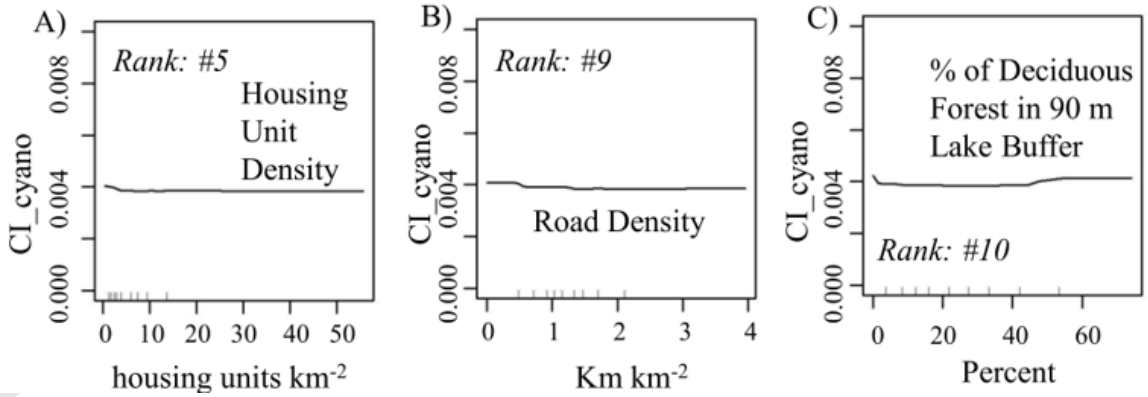


Fig. 6. Examples of partial distribution plots (PDPs) generated from the 100 random forest runs for "High Elevation" (n = 167) lakesheds.

In summary, CI_cyano distribution across the low elevation lake/reservoir stratum depicts a predominance of lakes/reservoirs within the 'high' cyanobacteria cell count level WHO guidance/action level (WHO, 2003). This trend was expected given the anthropogenic influence on the LC present within these lakesheds. Seven of the top ten variables within this designation were related to agricultural practices and landscapes. The high elevation lakes/reservoirs designation dominated the moderate to low WHO cyanobacteria cell count levels, again based on

the higher percentage of natural vegetation present within these lakesheds (Fig. 7). Model comparisons showed that the high elevation CI_cyano model resulted in 3 times lower model error when compared to the low elevation CI_cyano model. This was unexpected given the better fitting correlation of the top 10 VOI ranks to CI_cyano with the lower elevation lakes/reservoirs (Table 4).

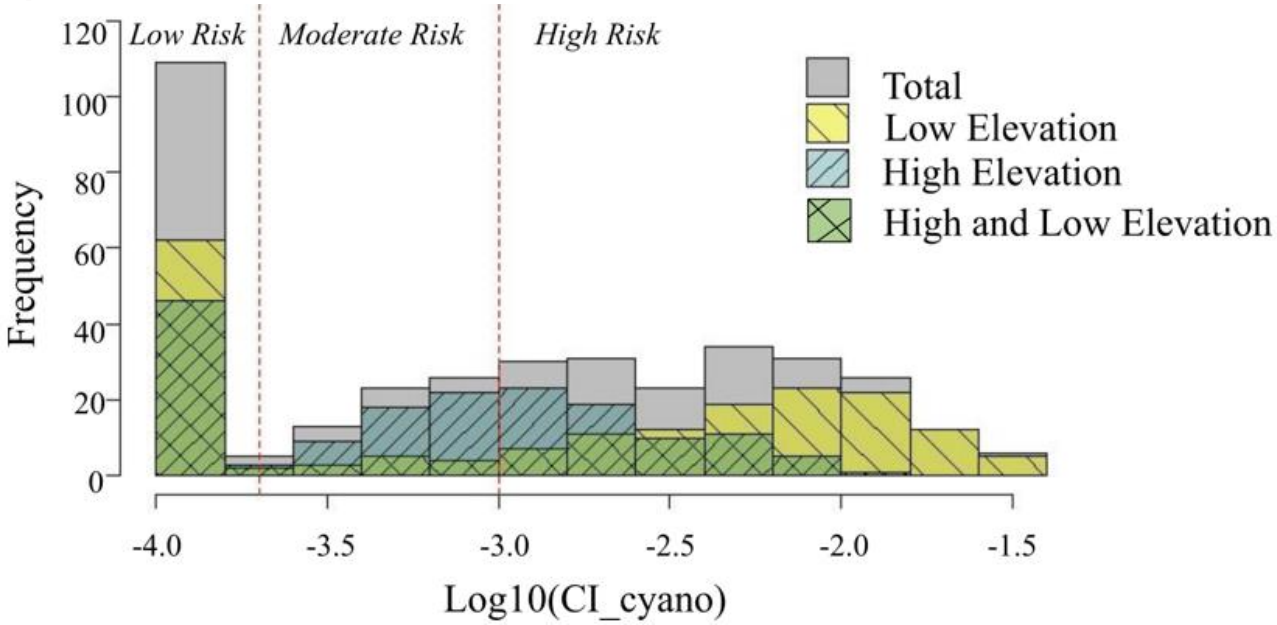


Fig. 7. Histogram of CI_cyano distributions within both low and high elevation lake/reservoir designations within the study area. Cyanobacteria risk thresholds for low (<20,000 cells/mL), moderate (20,000 – 100,000 cells/mL), and high (>100,000 – 1,000,000 cells/mL) are described in the World Health Organization (2003) recreational guidance/action levels for cyanobacteria, chlorophyll a, and microcystin (WHO, 2003). Note: Cyano Abundance (cells/mL) = CI_cyano * 10^8 (Lunetta et al., 2015).

4.3. CART Analysis

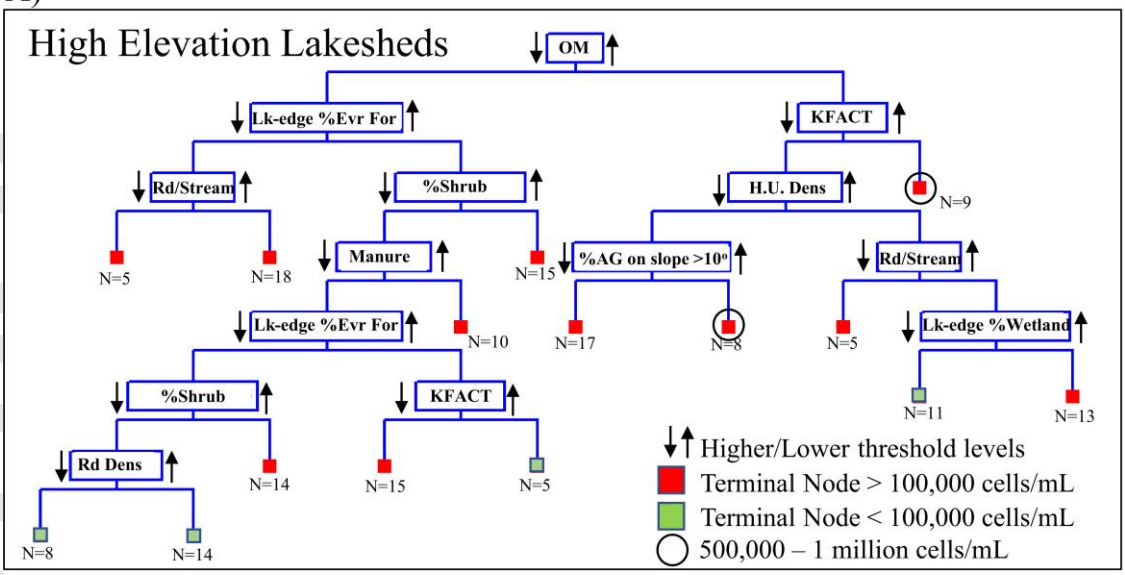
CART analysis is a tool that can assist water quality managers to geospatially identify drivers in CI_cyano development with the possibility of mitigating those inputs to reduce impact on these freshwater systems. CART for the high elevation lakeshed analysis identified organic matter at the root node with further splits resulting in 15 terminal nodes, of which all but four (38 of 167 lakes) exceeded the 100,000 cells/mL WHO threshold for high cyanobacteria risk (Fig. 8A). Of the 11 terminal nodes that exceeded the 100,000 cells/mL threshold, only two exceeded 500,000 cells/mL. In contrast, the low elevation lakeshed CART analysis resulted in 16 terminal nodes of which 14 exceeded the 100,000 cells/mL. Of these 14, nine exceeded 500,000 cells/mL with five of these exceeding 1,000,000 cells/mL (Fig. 8B). In application we display the lakes of the CART branch that track the highest CI_cyano (> 1,000,000 cells/mL) for low elevation designated lakesheds in Minnesota (Fig. 8B, 9). Here we see that even small percentages of artificial drainage existent within a lakeshed drive CI_cyano to higher levels compared to when

tile drainage is absent. Decision makers may consider managing drivers identified within these regression trees to possibly mitigate CI_cyano development within these lakes/reservoirs.

The high/low elevation lake/reservoir designations present differing cyanoHAB drivers as visualized by comparing the agricultural/urban dominated landscape (Lake Mendota, Wisconsin) to a forested/wetland dominated landscape (Perch Lake, Michigan) (Fig. 10). Within these two very different systems CI_cyano falls in the WHO 'high' range (Perch Lake: 197,789 cells/mL, Lake Mendota: 223,872 cells/mL). Within the high elevation designation, Perch Lake (46.3614°, +88.6583°), a 4.22 km^2 lake at a 4.3 meter maximum depth, located in Iron County, Michigan, is dominated by forest (58.4%) and wetland (38.5%), with no urban component. In 2018 the Cooperative Lakes Monitoring Program rated this lake between the mesotrophic (mid-ranged nutrient amounts) and eutrophic classification (slightly more than mesotrophic). This lake is characterized by organic matter greater than 12.38% and a soil erosion factor greater than 0.20. The thresholds for both these variables resulted in this lake categorized in a terminal node with a high WHO rating for cyanobacteria risk.

Despite a similar cyanobacteria cell count, the low elevation designated Lake Mendota (43.0949°, -89.3701°), a 39.8 km^2 lake situated in Dane County, Wisconsin with a maximum depth of 25.3 meters, was influenced by both anthropomorphic activity, lake physiography, and meteorological events. The lakeshed is dominated by an agricultural (69.6%) and urban (16.0%) landscape. The Wisconsin Department of Natural Resources has rated this lake as eutrophic since 2016. This lake is characterized by seven splits down one branch in the CART tree with the terminal node determined by thresholds from these variables: tile drained agricultural land, evergreen land cover, ammonia-N application rate, labile loss of P due to runoff, runoff, mean lake depth, and sediment loading (Fig. 8B). Of these possible drivers, watershed managers may possibly mitigate nutrient inputs (N and P) and/or establish land cover practices that reduce sediment loading into this system. Of interest, in 2009 Lake Mendota was invaded by a Ponto-Caspian daphnid called *Bythotrephes longimanus*, or the spiny water flea. This zooplankton is a voracious consumer of grazers, and recent studies have suggested that a great deal of the increase in phytoplankton biomass can be attributed to the loss of the grazing guild due to *B. longimanus* predation (Walsh et al., 2016). This may indicate a biological driver not captured in our models.

A)



B)

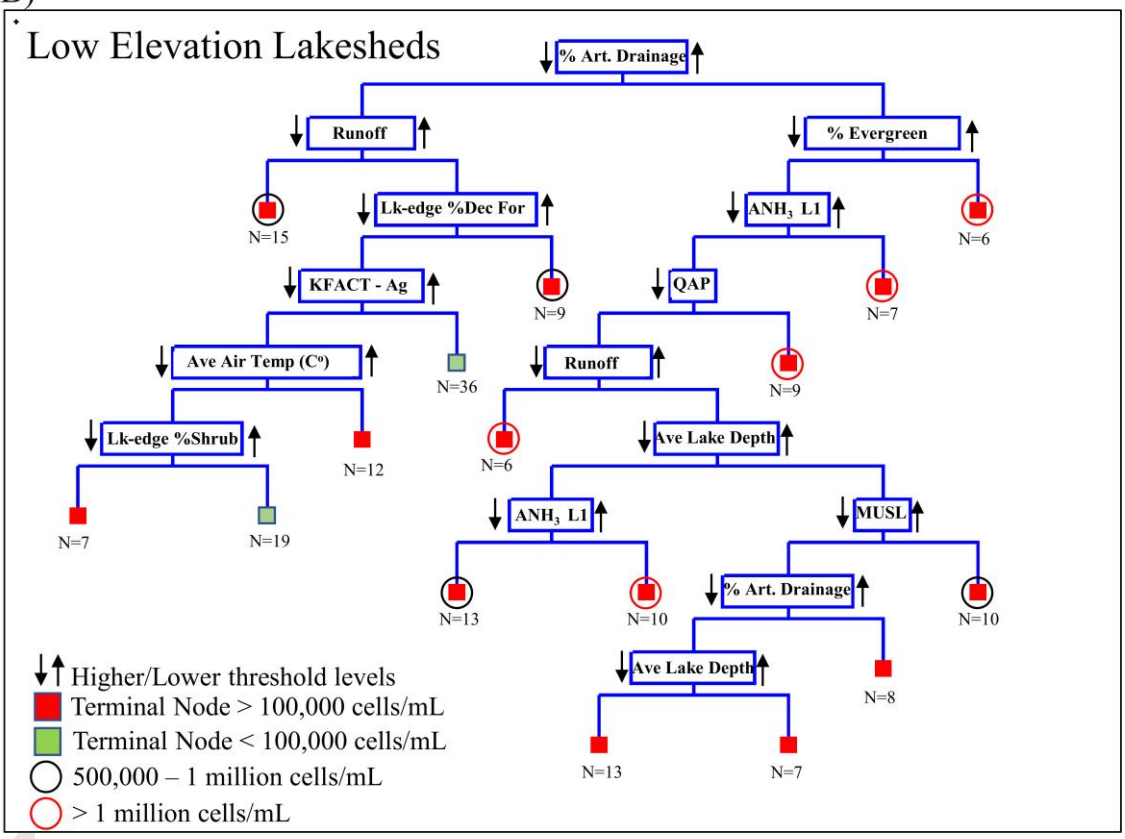


Fig. 8. CART representation for both low and high elevation lakeshed designations. Note: Lk – lake, Art. – artificial, ANH_3 L1 – surface NH_3 application, Dec. – deciduous forest, MUSL – sediment, QAP – labile P loss in runoff, KFACT – soil erodibility, Rd/Stream – Road-stream intersection density, H.U. Dens. – housing unit density, OM – organic matter, Rd Dens – road density.

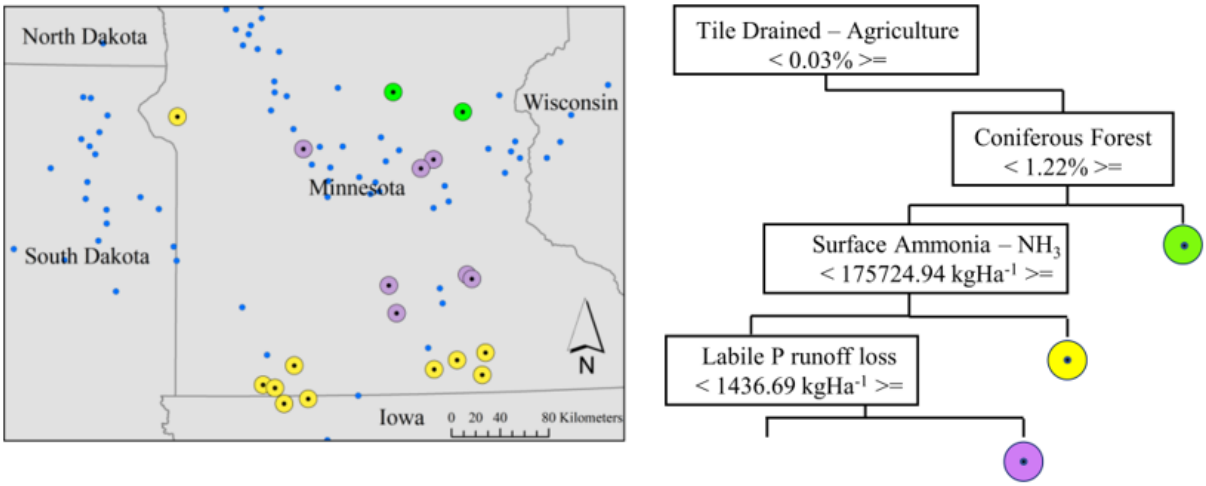


Fig. 9. An example of following one branch of this low elevation designation tree where each terminal node resulted in the "high' WHO range (100,000 – 1,000,000 cells/mL). Any tile drainage drives CI_cyano values to the higher levels. On this branch, lakesheds with a presence of coniferous forest, with higher surface N (ammonia) applications and with labile P loss from runoff are mapped. Note: other low elevation designated lakes are depicted with smaller solid blue circles.

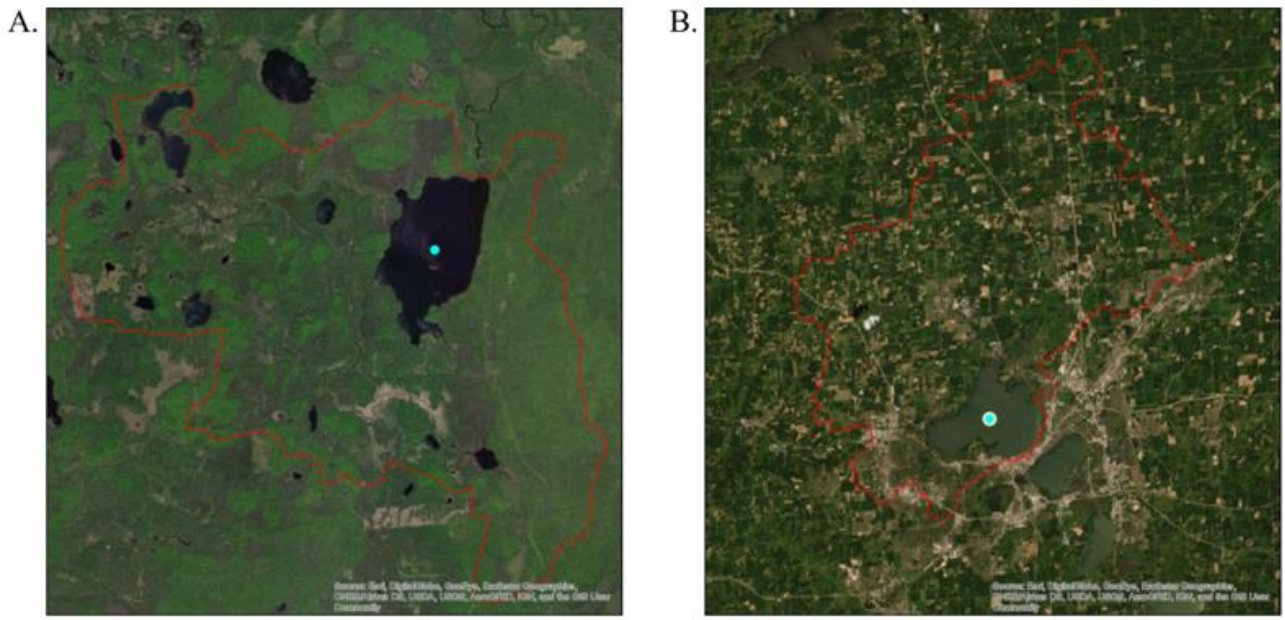


Fig. 10. Perch Lake, Michigan (A) – a high elevation designation and Lake Mendota, Wisconsin (B) – a low elevation designation, both with the highest modelled terminal node CI_cyano's for their stratifications.

5 Conclusions

In this study we have identified a suite of VOIs across two landscape strata: high versus low elevation lakes in the north-central United States. We also investigated the hierarchical spatial structure in the relationships between these VOIs that reflect the condition of grouped lakes. In summary, CI_cyano was quite notably higher in the low elevation lakes when compared to the high elevation lakes (Fig. 7, Table 2). Six of the top ten drivers identified across all 369 lakes were agrarian-related (i.e., nutrient application, artificial drainage, etc.). Eight of the top ten variables for the low elevation lakes were positively correlated with CI_cyano, a trend opposite that found with the high elevation lakes. Air temperature, the surrogate for lake temperature, was ranked in the top 25 VOI rankings for only the high elevation lake designation yet was not highly ranked (VOI rank = 19). This result was surprising given the prominence of cyanobacteria development related to increased temperatures in many previous studies.

There are high complexities involved in the identification of possible drivers in cyanoHAB development; drivers that both may impede or enhance these bloom occurrences (Richardson et al., 2018). Prior studies have been specific to certain systems – challenging the applicability beyond the ecosystem studied. With the use of CI_cyano derived from the MERIS sensor, we were able to expand the sample size to a large geographic range over the north-central United States. Regardless of the landscape, whether dominated by anthropogenic or environmental influences, cyanoHABs may develop at a similar intensity and recurrence. This suggests that there are concurrent interactions that require further investigation.

Acknowledgements

This work was supported by the NASA Ocean Biology and Biogeochemistry Program/Applied Sciences Program (proposal 14-SMDUNSOL14-0001) and by U.S. EPA, NOAA, U.S. Geological Survey Toxic Substances Hydrology Program, and Oak Ridge Institute for Science and Technology (ORISE). The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the U.S. EPA. Mention of trade names or commercial products does not constitute endorsement or recommendation for use by the U.S. Government. R-code and lakeshed data can be accessed at: https://edg.epa.gov/metadata/catalog/main/home.page. We are grateful for the LakeCat data shared by EPA Western Ecology Division ORD scientists, Ryan Hill and Scott Leibowitz. Donald Ebert also aided with riparian geodata processing. EPIC data was processed by Ellen Cooter. Jeff Hollister provided lake metrics (lakemorpho) for this study. We thank the anonymous reviewers for their comments and suggestions.

References

Anderson, D.M., Glibert, P.M., & Burkholder, J.M. (2002). Harmful algal blooms and eutrophication: nutrient sources, composition, and consequences. *Estuaries, 25*, 704–726. https://doi.org/10.1007/BF02804901

Anderson, C.R., Moore, S.K., Tomlinson, M.C., Silke, J., & Cusack, C.K. (2015). Living with harmful algal blooms in a changing world: strategies for modeling and mitigating their effects in coastal marine ecosystems. In: Shroder, J.F., Ellis, J.T., Sherman, D.J.(Eds.), Coastal and Marine Hazards, Risks, and Disasters, pp. 495–561. https://doi.org/10.1016/B978-0-12-396483-0.00017-0

Baker, M.E., Weller, D.E., & Jordan, T.E. (2006). Improved methods for quantifying potential nutrient interception by riparian buffers. *Landscape Ecology, 21*, 1327–1345. https://doi.org/10.1007/s10980-006-0020-0

Beaver, J.R., Manis, E.E., Loftin, K.A., Graham, J.L., Pollard, A.I., & Mitchell, R.M. (2014). Land use patterns, ecoregion, and microcystin relationships in US lakes and reservoirs: a preliminary evaluation. *Harmful Algae, 36*, 57–62. http://dx.doi.org/10.1016/j.ecoleng.2017.07.032

Bell, C.D., Tague, C.L., & McMillan, S.K. (2019). Modeling runoff and nitrogen loads from a watershed at different levels of impervious surface coverage and connectivity to storm water control measures. *Water Resources Research, 55*, 2690– 2707. https://doi.org/10.1029/2018WR023006

Bentrup, G.M., (2008). Conservation Buffers: Design Guidelines for Buffers, Corridors, and Greenways. General Technical Reports SRS- 109. Department of Agriculture, Forest Service, Southern Research Station, Asheville, North Carolina, 110 pp. Available online at http://www.srs.fs.usda.gov/pubs/gtr/gtr_srs109.pdf. Accessed [06/11/2021].

Boström, B., Pettersson, A.K., & Ahlgren, I. (1989). Seasonal dynamics of a cyanobacteria-dominated microbial community in surface sediments of a shallow, eutrophic lake. *Aquatic Science 51*, 153–178. https://doi.org/10.1007/BF00879300

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Chirico, N., António, D.C., Pozzoli, L., Marinov, D., Malagó, A., Sanseverino, I., Beghi, A., Genoni, P., Dobricic, S., & Lettieri, T. (2020). Cyanobacterial Blooms in Lake Varese: Analysis and Characterization over Ten Years of Observations. *Water, 12*, 675. https://doi.org/10.3390/w12030675

Christensen, J.R., Nash, M.S., & Neale, A. (2013). Identifying riparian buffer effects on stream nitrogen in southeastern coastal plain watersheds. *Environmental Management, 52*, 1161–1176. DOI:10.1007/s00267-013-0151-4

Clark, J.M., Schaeffer, B.A., Darling, J.A., Urquhart, E.A., Johnston, J.M., Ignatius, A.R., Myer, M.H., Loftin, K.A., Werdell, P.J., & Stumpf, R.P. (2017). Satellite monitoring of cyanobacterial harmful algal bloom frequency in recreational waters and drinking water sources. *Ecological Indicators, 80*, 84–95. https://doi.org/10.1016/j.ecolind.2017.04.046

Clover, M.W. (2005). Impact of nitrogen management on corn grain yield and nitrogen loss on a tile drained field, (M. S. Thesis). Retrieved from U. of Illinois at Urbana-Champaign. (https://i-share.carli.illinois.edu/vf-uiu/Record/UIUdb.5178317/Holdings#tabnav). Urbana, IL: University of Illinois.

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
Coffer, M.M., Schaeffer, B.A., Salls, W.B., Urquhart, E., Loftin, K.A., Stumpf, R.P., Werdell, P.J., & Darling J.A. (2021). Satellite remote sensing to assess cyanobacterial bloom frequency across the United States at multiple spatial scales. *Ecological Indicators, 128*, 107822. https://doi.org/10.1016/j.ecolind.2021.107822

Conley, D.J., Paerl, H.W., Howarth, R.W., Boesch, D.F., Seitzinger, S.P., Havens, K.E., Lancelot, C., & Likens, G.E. 2009. Controlling eutrophication: nitrogen and phosphorus. *Science 323*, 1014–15.

Downing, J.A., Watson, S., & McCauley, E. (2001). Predicting Cyanobacteria dominance in lakes. *Canadian Journal of Fisheries and Aquatic Sciences, 58*, 1905–1908. https://doi.org/10.1139/f01-143

Drury, C.F., Tan, C.S., Gaynor, J.D., Oloya, T.O., & Welacky, T.W. (1996). Influence of controlled drainage-subirrigation on surface and tile drainage nitrate loss. *Journal of Environmental Quality, 25*, 317. https://doi.org/10.2134/jeq1996.00472425002500020016x

Duan, H., Ma, R., Xu, X., Kong, F., Zhang, S., Kong, W., Hao, J., & Shang, L. (2009). Two-decade reconstruction of algal blooms in China's Lake Taihu. *Environmental Science and Technology, 43*, 3522–3528. https://doi.org/10.1021/es8031852

Ebert, D.W. & Wade, T.G. (2004). Analytical Tools interface for Landscape Assessments (ATtILA) Version 2004 User Manual. *US-EPA Report*, EPA/600/R-04/083.

Elser, J.J., Bracken, M.E.S., Cleland, E.E., Gruner, D.S., Harpole, W.S., Hillebrand, H., Bgai, J.T., Seabloom, E.W., Shurin, J.B., Smith, J.E. (2007). Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. *Ecololgy Letters 10*, 1124e1134. https://doi.org/10.1111/j.1461-0248.2007.01113.x

Falconer, I.R. & Humpage, A.R. (1996). Tumour promotion by cyanobacterial toxins. *Phycologia, 35*, 74–79. https://doi.org/10.2216/i0031-8884-35-6S-74.1

Fraterrigo, J.M., & Downing, J. (2008). The Influence of land use on lake nutrients varies with watershed transport capacity. *Ecosystems, 11*, 1021−1034. https://doi.org/10.1007/s10021-008-9176-6

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189–1232.

Garcia-Pichel, F., & Wojciechowski, M.F. (2009). The Evolution of a Capacity to Build Supra-Cellular Ropes Enabled Filamentous Cyanobacteria to Colonize Highly Erodible Substrates. *PLOS ONE, 4*, 11, e7801. https://doi.org/10.1371/journal.pone.0007801

Graham, J.L., Dubrovsky, N.M. & Eberts, S.M. (2016). Cyanobacterial harmful algal blooms and U.S Geological Survey science capabilities. United States Geological Survey. Available online at https://pubs.er.usgs.gov/publication/ofr20161174. Accessed [06/11/2021].
Hayes, N. & Vanni, M. (2018). Microcystin concentrations can be predicted with phytoplankton biomass and watershed morphology. *Inland Waters, 8*, 1–11. https://doi.org/10.1080/20442041.2018.1446408

Heisler, J., Glibert, P.M., Anderson, D.M., Cochlan, W., Dennison, W.C., Dortch, Q., Gobler, C. J., Heil, C.A., Humphries, E., Lewitus, A., Magnien, R., Marshall, H.G., Sellner, K., Stockwell, D.A., Stoecker, D.K., & Suddleson, M. (2008). Eutrophication and harmful algal blooms: A scientific consensus. *Harmful Algae, 8*, 103–110. https://doi.org/10.1016/j.hal.2008.08.006

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

Hollister, J.W., & Stachelek, J.J. (2017). lakemorpho: Calculating lake morphometry metrics in R. *F1000Research, 6*, 1718. https://doi.org/10.12688/f1000research.12512.1

Kleinman, P.J.A., Srinivasan, M.S., 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 Environment Quality, 35*, 1248–1259. https://doi.org/10.2134/jeq2006.0015

Knapp, A.S. & Milewski, A.M. (2020). Spatiotemporal Relationships of Phytoplankton Blooms, Drought, and Rainstorms in Freshwater Reservoirs. *Water*, *12*, 404. https://doi.org/10.3390/w12020404

Lawrence, R.L. & Wright, A. (2001). Rule-based classification systems using classification and regression tree (CART) analysis. *Photogrammetric Engineering Remote Sensing, 67*, 1137–1142.

Lee, J. J., (2014). Driven by Climate Change, Algae Blooms Behind Ohio Water Scare Are New Normal. *Nat. Geo*, August 6, 2014.

Likens, G. E. (1972). *Nutrients and eutrophication*. American Society of Limnology *Oceanography special symposium, 1*, 328.

Liu, W., Zhang, Q., & Liu, G. (2011). Effects of watershed land use and lake morphometry on trophic state of Chinese lakes: implications for eutrophication control. *Clean—Soil, Air, Water, 39*, 35–42. https://doi.org/10.1002/clen.201000052

Lunetta, R.S., Schaeffer, B.A., Stumpf, R.P., Keith, D., Jacobs, S.A., & Murphy, M.S. (2015). Evaluation of cyanobacteria cell count detection derived from MERIS imagery across the eastern USA. *Remote Sensing of Environment, 157,* 24–34. https://doi.org/10.1016/j.rse.2014.06.008

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2013). Cluster analysis basics and extensions, R Package version 1.14.4.

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

McKay, L., Bondelid, T., Dewald, T., Johnston, J., Moore, R., & Rea, A. (2012). NHDPlus Version 2: User Guide. National Operational Hydrologic Remote Sensing Center, Washington, D.C.

Medina-Cobo, M., Domínguez J.A., Quesada A., & Hoyos Cd. (2014). Estimation of cyanobacteria biovolume in water reservoirs by MERIS sensor. *Water Research, 63*, 10–20. https://doi.org/10.1016/j.watres.2014.06.001

Michalak, A.M., Anderson, E.J., Beletsky, D., Boland, S., Bosch, N.S., Bridgeman, T.B., Chaffin, J.D., Cho, K., Confesor, R., Daloğlu, I., DePinto, J.V. (2013). Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions. *Proceedings of the National Academy of Sciences*, *110*, 6448–6452. DOI: 10.1073/pnas.1216006110

Michaud, A.R., Poirier, S.-C., & Whalen, J.K. (2019). Tile Drainage as a Hydrologic Pathway for Phosphorus Export from an Agricultural Subwatershed. *J. Environ. Qual., 48*, 64–72. https://doi.org/10.2134/jeq2018.03.0104

Miller, K., & Sangalang, J. (2021). Toxin in West Palm Beach drinking water: Do not boil tap water, things to know. In The Palm Beach Post, Available online at https://www.palmbeachpost.com/story/news/2021/06/01/toxic-water-west-palm-beach-ok-drink-what-can-happen/5289411001/. Accessed [06/11/2021].

Mishra, S., Stumpf, R.P., Schaeffer, B., Werdell, P.J., Loftin, K.A., & Meredith, A. (2021). Evaluation of a satellite-based cyanobacteria bloom detection algorithm using field-measured microcystin data. *Science of the Total Environment, 774*, 145462. https://doi.org/10.1016/j.scitotenv.2021.145462

Moore, D.S., Notz, W.I., & Fligner, M.A. (2018). Chapter 4. In *The basic practice of statistics* (Chapter 4: "Scatterplots and Correlation"). New York, NY.

Morford, S.L., Houlton B.Z., & Dahlgren R.A. (2011). Increased forest ecosystem carbon and nitrogen storage from nitrogen rich bedrock. *Nature, 477*, 78–81. https://doi.org/10.1038/nature10415

Mrdjen, I., Fennessy, S., Schaal, A., Dennis, R., Slonczewski, J.L., Lee, S., & Lee, J. (2018). Tile drainage and anthropogenic land use contribute to harmful algal blooms and microbiota shifts in inland water bodies. *Environmental Science and Technology, 52*, 8215–8223. https://doi.org/10.1021/acs.est.8b03269

Omernik, J.M. (1987). Ecoregions of the conterminous United States. Map (scale 1:7,500,000). *Annals of the Association of American Geographers, 77*, 118–125.

Omernik, J.M. (1995). Ecoregions: A spatial framework for environmental management. In: *Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making* (pp. 49–62). Boca Raton, FL.

Padilla, A. 2018. Landscape and climate drivers of harmful algal blooms in Iowa (thesis). Retrieved from Electronic Theses and Dissertations (https://scholarworks.uni.edu/etd/523). Cedar Falls, IA: University of Northern Iowa.

Paerl, H.W., & Huisman, J. (2008). Blooms like it hot. Science. 320, 57–58. https:// 10.1126/science.1155398

Paerl, H.W. (2008). Nutrient and other environmental controls of harmful cyanobacterial blooms along the freshwater-marine continuum. *Advances in Experimental Medicine and Biology, 619*, 216–241. https://doi.org/10.1007/978-0-387-75865-7_10

Paerl, H.W., Xu, H., McCarthy, M.J., Zhu, G., Qin, B., Li, Y., & Gardner, W.S. (2011). Controlling harmful cyanobacterial blooms in a hyper-eutrophic lake (Lake Taihu, China): The need for a dual nutrient (N & P) management strategy. *Water Research, 45*, 1973–83. https://doi.org/10.1016/j.watres.2010.09.018

Paerl, H.W., & Paul, V.J. (2012). Climate change: Links to global expansion of harmful cyanobacteria. *Water Research, 46*, 1349–1363. https://doi.org/10.1016/j.watres.2011.08.002

Paerl, H.W. & Otten, T.G. (2013). Harmful cyanobacterial blooms: Causes, consequences, and controls. *Microbial Ecology, 65*, 995–1010. https://doi.org/10.1007/s00248-012-0159-y

Prasad, A., Iverson, L., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems, 9*, 181–199.
PRISM Climate Group (2012). PRISM climate data. Available online at http://prism.oregonstate.edu. Accessed [06/11/2021].

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at http://www.R-project.org/. Accessed [06/11/2021].

Reichwaldt, E.S. & Ghadouani, A. (2012). Effects of rainfall patterns on toxic cyanobacterial blooms in a changing climate: between simplistic scenarios and complex dynamics. *Water Research, 46*, 1372–1393. https://doi.org/10.1016/j.watres.2011.11.052

Remson, I., Randolph, J.R., & Barksdale, H.C. (1959). Zone of Aeration and Its Relationship to Ground Water Recharge. *American Water Works Association, 51*, 3, 371–378.

Richardson, J., Miller, C., Maberly, S.C., Taylor, P., Globevnik, L., Hunter, P., Jeppensen, U., Moe, J., Pasztaleniec, A., Søndergaard, M. & Carvalho, L. (2018). Effects of multiple stressors on cyanobacteria abundance vary with lake type. *Glob Change Biol., 24*, 5044– 5055. https://doi.org/10.1111/gcb.14396.

Schaeffer, B.A., Bailey, S.W., Conmy, R.N., Galvin, M., Ignatius, A.R., Johnston, J.M., Keith, D.J., Lunetta, R.S., Parmar, R., Stumpf, R.P., Urquhart, E.A., Werdell, P.J., & Wolfe, K. (2018). Mobile device application for monitoring cyanobacteria harmful algal blooms using Sentinel-3 satellite Ocean and Land Colour Instruments. *Environmental Modelling Software, 109*, 93 –103. https://doi.org/10.1016/j.envsoft.2018.08.015

Schindler, D.W. (1975). Whole-lake eutrophication experiments with phosphorus, nitrogen and carbon. *Verhandlungen der Internationalen Vereinigung fur Theoretische und Angewandte Limnologie, 19*, 3221–3231.

Schindler, D.W. (1977). The evolution of phosphorus limitation in lakes. *Science, 195*, 260–262.

Sivonen, K. (1996). Cyanobacterial toxins and toxin production. *Phyclogia, 35*,12–24.

Smith, D.R., King, K. W., Johnson, L., Francesconi, W., Richards, P., Baker, D. & Sharply, A.N. (2015). Surface runoff and tile drainage transport of phosphorus in the midwestern United States. *Journal of Environmental Quality, 44*, 495−502. doi:10.2134/jeq2014.04.0176

Soil Survey Division Staff (1993). Soil survey manual. Soil Conservation Service. U.S. Department of Agriculture Handbook 18. Chapter 3. p.31.

Song, K., Li, L., Li, Z., Tedesco, L.P., Hall, B.E., & Shi, K. (2013). Remote detection of cyanobacteria through phycocyanin for water supply source using three-band model. *Ecological Informatics, 15*, 22–33. https://doi.org/10.1016/j.ecoinf.2013.02.006

Song, K., Li, L., Tedesco, L.P., Li, S., Hall, B.E., & Du, J. (2014).
Remote quantification of phycocyanin in potable water sources through an adaptive model. *ISPRS J. Photogramm. Remote Sens., 95*, 68–80. https://doi.org/10.1016/j.isprsjprs.2014.06.008

Soranno, P.A., Cheruvelil, K.S., Wagner, T., Webster, K.E., & Bremigan, M.T. (2015). Effects of land use on lake nutrients: the importance of scale, hydrologic connectivity, and region. *PLoS ONE, 10*: e0135454. https://doi.org/10.1371/journal.pone.0135454

Steinberg, D., & Golovnya, M. (2006). CART 6.0 Users's Manual. San Diego, CA: Salford Systems.

Strobl, C., Boulesteix, A.L., Zeileis, A., & Hothorn, T. (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics, 8,* 25. https://doi.org/10.1186/1471-2105-8-25

Strobl, C., Boulesteix, A-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9,* 307. https://doi.org/10.1186/1471-2105-9-307.

Strobl, C. & Zeileis, A. (2008). Danger: High Power! – Exploring the Statistical Properties of a Test for Random Forest Variable Importance. *Department of Statistics, Ludwig Maximilian University,* Technical Reports, No.17., Munich, Germany.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests *Psychological Methods, 14* (4), 323–348. https://doi.org/10.1037/a0016973

Sugg, Z. (2007). Assessing US farm drainage: Can GIS lead to better estimates of subsurface drainage extent? World Resources Institute, Washington D.C., 1–8, August.

Taranu, Z.E., Gregory-Eaves, I., Steele, R.J., Beaulieu, M., & Legendre, P. (2017). Predicting microcystin concentrations in lakes and reservoirs at a continental scale: a new framework for modelling an important health risk factor. *Global Ecology and Biogeography, 00,* 1–13. http://dx.doi.org/10.1111/geb.12569

The Nowak Consulting Group (2018). City of Salem Water Advisory After-Action Assessment. Available online at https://www.cityofsalem.net/CityDocuments/water-advisory-after-action-report-2018.pdf. Accessed [06/11/2021].

Toffolon, M., Piccolroaz, S., Majone, B., Soja, A-M., Peeters, F., Schmid M., & West, A. (2014). Prediction of surface temperature in lakes with different morphology using air temperature Limnol. *Oceanogr., 59,* 2182–202. https://doi.org/10.4319/lo.2014.59.6.2185

Tomaschek, F., Hendrix, P., & Baayen, R.H. 2018. Strategies for managing collinearity in multivariate linguistic data. *Journal of Phonetics, 71,* 249–267. https://doi.org/10.1016/j.wocn.2018.09.004

U.S. Department of Agriculture (2006). US General Soil Map (STATSGO). National Geospatial Center of Excellence, Natural Resource Conservation Service, US Department of Agriculture, Washington, DC. (Available online at http://www.ncgc.nrcs.usda.gov/products/datasets/statsgo/. Accessed [06/11/2021].

U.S. Environmental Protection Agency (2010). National Aquatic Resource Surveys. National Lakes Assessment 2007 (data and metadata files). Available online at http://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys. Accessed [06/11/2021].

Vanderhoof, M.K., Christensen, J.R., & Alexander, L.C. (2017). Patterns and drivers for wetland connections in the Prairie Pothole Region, United States. *Wetlands Ecological Management, 25,* 275–297. https://doi.org/10.1007/s11273-016-9516-9

Van Esbroeck, C.J., Macrae, M.L., Brunke, R.I. & McKague, K. (2016). Annual and seasonal phosphorus export in surface runoff and tile drainage from agricultural fields with cold temperate climates. *Journal of Great Lakes Research, 42*, 1271−1280. https://doi.org/10.1016/j.jglr.2015.12.014

Walsh, J.R., Stephen, R., Carpenter, M., & Vander Zanden, J. (2016). Invader triggers loss of ecosystem service. *Proceedings of the National Academy of Sciences, 113*(15), 4081-4085; DOI: 10.1073/pnas.1600366113

Wang, M., & Shi, W. (2008). Satellite-observed algae blooms in China's Lake Taihu. *Eos, 89,* 201–202. https://doi.org/10.1029/2008EO220001.

Weihrauch, C., & Opp, C. (2018). Ecologically relevant phosphorus pools in soils and their dynamics: the story so far. *Geoderma, 325*, 183–194. https://doi.org/10.1016/j.geoderma.2018.02.047

Weller, D.E., Baker, M.A., & Jordan, T.E. (2011). Effects of riparian buffers on nitrate concentrations in watershed discharges: new models and management implications. *Ecological Applications, 21*,1679–1695. https://doi.org/10.1890/10-0789.1

WHO (2003). Guidelines for Safe Recreational Water Environments: Volume 1: Coastal and Freshwaters. World Health Organization.

Williams, J.W., Izaurralde, R.C., & Steglich, E.M. (2012). Agricultural Policy/Environmental eXtender Model Theoretical Documentation Version 0806; Blackland Research and Extension Center: Temple, TX, USA, 1–131.

Wynne, T.T., Stumpf, R.P., Tomlinson, M.C., Warner, R.A., Tester, P.A., Dyble, J., & Fahnenstiel, G.L. (2008). Relating spectral shape to cyanobacterial blooms in the Laurentian Great Lakes. *International Journal of Remote Sensing, 29*, 3665–3672. https://doi.org/10.1080/01431160802007640

Yang, L., Jin, S., Danielson, P., Homer, C., Gass, L., Case, A., Costello, C., Dewitz, J., Fry, J., Funk, M., Grannemann, B., Rigge, M., & Xian, G. (2018). A new generation of the United States National Land Cover Database: requirements, research priorities, design, and implementation strategies. *ISPRS Journal of Photogrammetry and Remote Sensing, 146*, 108–123. https://doi.org/10.1016/j.isprsjprs.2018.09.006

Yuan, Y., Wang, R., Cooter, E., Ran, L., Daggupati, P., Yang, D., Srinivasan, R., & Jalowska, A. (2018). Integrating multimedia models to assess nitrogen losses from the Mississippi River basin to the Gulf of Mexico. *Biogeosciences, 15,* 7059–7076. https://doi.org/10.5194/bg-15-7059-2018. https://doi.org/10.5194/bg-15-7059-2018