UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

SPATIAL AND TEMPORAL PERSPECTIVES ON CYANOBACTERIAL ECOLOGICAL DYNAMICS USING SPACE-BASED AND GENOMIC APPROACHES

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

KATHERINE V. COOK
Norman, Oklahoma
2023

SPATIAL AND TEMPORAL PERSPECTIVES ON CYANOBACTERIAL ECOLOGICAL DYNAMICS USING SPACE-BASED AND GENOMIC APPROACHES

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF BIOLOGY

BY THE COMMITTEE CONSISTING OF

Dr. K. David Hambright, Chair

Dr. Katharine Marske

Dr. Xiangming Xiao

Dr. Amy McGovern

ACKNOWLEDGEMENTS

I have been incredibly lucky to have met and been supported by so many amazing people during my time at OU. I am indebted to my advisor, Dr. K. David Hambright, for his mentorship and support throughout my degree. I would like to thank Dr. Jessica Beyer, for always being willing to give advice or grab a coffee. I would like to thank the members of the Plankton Ecology and Limnology Lab, past and present, who helped with many aspects of these projects, from field sampling to data processing and analysis, specifically, Thayer Hallidayschult, Chris McLimans, Haiyuan Cai, Dani Glidewell, Eve Minkin, Chris Acy, and Claire Grimmett. A special acknowledgment to the undergraduates I had the privilege of working with and who helped immensely with field work: Garrett Eakers, Lena Wilson, and Wyatt Despain. I would like to acknowledge my fellow graduate students and friends who have supported me throughout this process: Edna Fernandez-Figueroa, Savannah Mapes, Emily Kiehnau, Rebecca Prather, and Michelle Busch.

My parents, David and Peggy, my grandmother, Bobbie, and my sister and her husband, Kristin and Aaron, have provided immeasurable support and love, and I could not have made it this far without them. And I am beyond grateful for my husband, Stephen, who has been a pillar of support throughout my studies and was always willing to read my first drafts. I would also like to thank his family, Tammy, Fred, Sarah, and Steven, for welcoming and supporting me.

I am grateful to many people who have made my success as a scientist possible. These included the members of my doctoral committee, Dr. Katharine Marske, Dr. Xiangming Xiao, and Dr. Amy McGovern; my co-authors; and Department of Biology staff, K. Baker, G. Martin, and E. Cooley.

Funding was provided by the OU Biology Department's Adams Scholarship and the L.G. Hill Excellence Fund Award, the Graduate Student Senate Research and Travel Grants, a Graduate Research Assistantship from the Oklahoma Department of Wildlife Conservation, the Nancy L. Mergler Dissertation Completion Fellowship, and a NSF Graduate Research Fellowship.

DEDICATION

To Stephen, Pepper, Goose, and Yumi,

your love, support, and cuddles mean everything.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 2

CHAPTER 3

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

ABSTRACT

Toxin-producing cyanobacteria, a type of blue-green algae, rank high among the global problems facing aquatic systems. Under certain conditions, particularly elevated nutrient concentrations created by agricultural and urban runoff, as well as altered temperature regimes driven by changing climate, cyanobacteria can form ecosystem disrupting blooms. Their toxins produce chemicals that are harmful, not only to other aquatic organisms, but to humans, pets, and terrestrial wildlife that come into contact with cyanobacterial toxins in water bodies. As such, harmful cyanobacteria blooms threaten freshwaters used for recreation, fisheries, and drinking water. Given the intricate dependency of humanity on water, and the concentration of population centers surrounding lakes and reservoirs, it is imperative that we find new and innovative ways to detect, monitor, manage, and mitigate cyanobacteria blooms. My research addresses these needs in multiple, but complimentary ways, and considers scales that range from satellite data that provides a vantage of blooms from hundreds of kilometers away, to next generation sequencing data that provides a view of how bloom communities form at the genomic level.

In my first two chapters, I search for the existence of a cyanobacteria-bacteria interactome. Cyanobacteria have evolved very small genomes compared to eukaryotes, and thus trade a reduced set of metabolic functions for energy savings during replication. It has been hypothesized that cyanobacteria rely on other bacteria in the community to perform these "lost" metabolic functions. This interaction between the cyanobacterial-bacterial community is termed an interactome and understanding this relationship could help us create innovative biological mitigation strategies in the future.

In chapter one, as an initial test of the cyanobacteria-bacteria interactome hypothesis, I examined a globally occurring, colonial cyanobacterium, *Microcystis aeruginosa*, and its

associated bacteria, testing the hypothesis that if *Microcystis* depends on bacteria for certain metabolic functions, then geographically distinct *Microcystis* blooms should harbor similar bacterial microbiome communities and metabolic functions. While I found that the bacteria species were different across nine lakes, the metabolic functions contributed by the associated bacteria were similar across globally distributed lakes, thus supporting the cyanobacteria-interactome hypothesis.

In chapter two, I focused on the hypothesized interactome of a different bloom forming cyanobacteria, *Raphidiopsis.* Unlike *Microcystis*, which has a mucilaginous outer layer that harbors a community of bacteria in close physical proximity to its cells, *Raphidiopsis* is a filamentous cyanobacteria, forming distinct multicellular trichomes with no sheath. I hypothesized if *Raphidiopsis* selects for certain bacteria, this would result in greater similarity in bacterial microbiome community composition within bloom phases than across bloom phases in the same year. After examining three years of bloom data, I found that bacterial communities associated with *Raphidiopsis* blooms were more similar to other communities in the same phase than communities within the same year. These results point toward *Raphidiopsis* selection of associated bacteria and are a first step in providing evidence for the existence of a co-evolved interactome.

In my third chapter, I move from genomes to satellites to explore rapid and accurate ways of detecting and monitoring cyanobacteria blooms. Rigorous monitoring for harmful cyanobacteria blooms is necessary to manage risks to animal and human health, but is often expensive and time-consuming. Remote sensing using satellite- and ground-based platforms has the potential to allow frequent and cost-effective monitoring of cyanobacterial blooms on lakes. Here, I test the capacity of satellite- and ground-based sensors to accurately predict

cyanobacteria blooms in Oklahoma. Contrary to research on larger, natural lakes, I found Landsat satellite models performed poorly and were unable to accurately detect blooms in Oklahoma reservoirs. The ground-based models, on the other hand, were highly accurate at detecting blooms. My results indicate alternatives to satellites, like ground-based sensors, should be further explored as reliable and inexpensive tools for monitoring harmful cyanobacteria blooms on large numbers of small lakes and reservoirs.

Chapter 1 - The global *Microcystis* interactome

Published in *Limnology and Oceanography*,
https://doi.org/10.1002/lno.11361

Katherine V. Cook,[1,2] C. Li,[3], H. Cai[1], L.R. Krumholz,[3] K.D. Hambright,[1,2,*] H.W. Paerl,[4] M.M. Steffen,[5] A.E. Wilson,[6] M. Burford,[7] H.-P. Grossart,[8] D. P. Hamilton,[7,9] H. Jiang,[10] A. Sukenik,[11] D. Latour,[12] E.I. Meyer,[13] J. Padisák,[14] B. Qin,[10] R.M. Zamor,[15,a] and G. Zhu[10]

[1]Plankton Ecology and Limnology Laboratory and [2]Program in Ecology and Evolutionary Biology, Department of Biology, The University of Oklahoma, Norman, Oklahoma

[3]Department of Microbiology and Plant Biology and Institute for Energy and the Environment, The University of Oklahoma, Norman, Oklahoma

[4]Institute of Marine Sciences, The University of North Carolina at Chapel Hill, Morehead City, North Carolina

[5]Department of Biology, James Madison University, Harrisonburg, Virginia

[6]School of Fisheries, Aquaculture, and Aquatic Sciences, Auburn University, Auburn, Alabama

[7]Australian Rivers Institute, Griffith University, Nathan, Queensland, Australia

[8]Department of Experimental Limnology, Leibniz Institute for Freshwater Ecology and Inland Fisheries, Stechlin, and Institute for Biochemistry and Biology, Potsdam University, Potsdam, Germany

[9]Environmental Research Institute, University of Waikato, Waikato, New Zealand

[10]State Key Laboratory of Lake Science and Environment, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing, China

[11]Israel Oceanographic and Limnological Research, The Yigal Allon Kinneret Limnological Laboratory, Migdal, Israel

[12]LMGE, Laboratoire Microorganismes: Génome et Environnement, Clermont Université, Université Blaise Pascal, Aubière, France

[13] Institute for Evolution and Biodiversity, University of Münster, Münster, Germany

[14]Department of Limnology, Institute of Environmental Science, University of Pannonia, Veszprém, Hungary

[15]Grand River Dam Authority, Vinita, Oklahoma

[a]Present address: Department of Natural Sciences, Northeastern State University, Tahlequah, Oklahoma

Abstract

Bacteria play key roles in the function and diversity of aquatic systems, but aside from study of specific bloom systems, little is known about the diversity or biogeography of bacteria associated with harmful cyanobacterial blooms (cyanoHABs). CyanoHAB species are known to shape bacterial community composition and to rely on functions provided by the associated bacteria, leading to the hypothesized cyanoHAB interactome, a coevolved community of synergistic and interacting bacteria species, each necessary for the success of the others. Here, we surveyed the microbiome associated with *Microcystis aeruginosa* during blooms in 12 lakes spanning four continents as an initial test of the hypothesized *Microcystis* interactome. We predicted that microbiome composition and functional potential would be similar across blooms globally. Our results, as revealed by 16S rRNA sequence similarity, indicate that *M. aeruginosa* is cosmopolitan in lakes across a 280° longitudinal and 90° latitudinal gradient. The microbiome communities were represented by a wide range of operational taxonomic units and relative abundances. Highly abundant taxa were more related and shared across most sites and did not vary with geographic distance, thus, like *Microcystis*, revealing no evidence for dispersal limitation. High phylogenetic relatedness, both within and across lakes, indicates that microbiome bacteria with similar functional potential were associated with all blooms. While *Microcystis* and the microbiome bacteria shared many genes, whole-community metagenomic analysis revealed a suite of biochemical pathways that could be considered complementary. Our results demonstrate a high degree of similarity across global *Microcystis* blooms, thereby providing initial support for the hypothesized *Microcystis* interactome.

Introduction

Seasonally recurrent harmful algal blooms, particularly those of toxic cyanobacteria (cyanoHABs), are a global phenomenon of growing concern impacting water quality, ecosystem services, and human health associated with freshwater systems (Paerl and Otten 2013). Accelerating eutrophication and climate change (e.g., rising temperatures and shifting hydrological regimes) has resulted in the proliferation, intensification, and prolongation of cyanoHABs around the world (Carey et al. 2012; O'Neil et al. 2012; Paerl and Paul 2012; Mantzouki et al. 2018). Water quality in freshwater systems is intimately linked to anthropogenic activities. With rapidly expanding agricultural and urban development, as well as prolonged stratification periods due to global warming, many systems have become eutrophic or at risk of eutrophication. Although there is debate regarding the roles of specific nutrients in cyanoHAB dynamics (Conley et al. 2009; Paerl et al. 2011, 2016; Schindler 2012; Schindler et al. 2016), there is general agreement that increased nutrient inputs lead to increases in cyanobacterial biomass. CyanoHABs can alter ecosystem function by causing anoxia, depleting dissolved nutrients, and shifting zooplankton communities, all which alter carbon flows (Paerl and Otten 2013). Such blooms can produce hundreds of secondary metabolites, including hepatotoxins, neurotoxins, and dermatotoxic irritants, all of which can pose serious health threats to humans, livestock, and wildlife (Carmichael 2001; Huisman et al. 2018).

CyanoHABs, once thought to be homogeneous populations, are now known to be accompanied by a diverse suite of heterotrophic bacteria (Eiler and Bertilsson 2004; Steffen et al. 2012; Xu et al. 2018), which may play an important role in cyanobacterial bloom health and duration. First, Bell and Mitchell (1972) defined this potential interactive relationship between cyanobacteria and heterotrophic bacteria as the "phycosphere" (Paerl and Kellar 1978, 1979;

Paerl and Millie 1996). It is well known that cyanobacteria generate abundant dissolved organic carbon resources to the benefit of nearby heterotrophs that can subsequently return benefits to the cyanobacteria, including removal of reactive oxygen species, CO2 generation, and nutrient recycling (Dziallas and Grossart 2011, 2012; Steffen et al. 2012; Paerl and Otten 2013). Moreover, *Microcystis* has been shown to alter ambient environmental conditions by decreasing oxygen concentrations and light availability (Paerl and Otten 2016), as well as by altering CO2 and pH levels (Havens 2008), which are likely to affect nearby bacteria. Indeed, studies have found that cyanobacterial bloom species strongly impact bacterial community composition (e.g., Nodularia—Salomon et al. 2003, *Microcystis*—Li et al. 2011; Steiner et al. 2017). Specifically, heterotrophic bacteria can form close-knit aggregates with *Microcystis*, and studies have shown greater similarity between attached Bacteria and Archaea communities than between free-living assemblages during *Microcystis* blooms (Cai et al. 2014; Yang et al. 2017; Batista et al. 2018; Xu et al. 2018).

One reason for this close association could be that, like other Bacteria and Archaea (Swan et al. 2013; Giovannoni et al. 2014), Cyanobacteria have small genomes compared with eukaryotes (Herdman et al. 1979; Humbert et al. 2013). While this may be beneficial for rapid reproduction and evolution, it is not necessarily conducive for cyanoHAB formation. Genome reduction can lead to loss of functions (Giovannoni et al. 2014), but can also confer a selective advantage if the organism can obtain the lost function through a public good as described by Morris et al. (2012) in the Black Queen Hypothesis. This hypothesis suggests that natural selection can act on "leaky" functions where a public good is produced and available to the whole community. Coupled with selection toward smaller genomes to reduce replication-related fitness costs, some members of the community can receive metabolic products as public goods,

which are useful metabolites or other necessary resources that are leaked into the cell-external environment. With such products available extracellularly, these "leaky" functions become dispensable and once lost confer a selective advantage to that organism (Pande and Kost 2017). Garcia et al. (2015) proposed that this coevolved community of synergistic and interacting bacteria species was a bacterial community microbiome or "interactome," analogous to the microbiome concept described for humans (Human Microbiome Project Consortium 2012), soils (Fierer 2017), and coral reefs (Bourne et al. 2013). We hypothesize that the bacteria associated with *Microcystis* may be providing functions that help sustain it during blooms.

Given the small size of bacterial genomes and presumed ubiquity of *Microcystis aeruginosa* (Kützing) Kützing, if there exists a mutualistic interactome, we would predict a microbiome of a certain species of associated bacteria or perhaps metabolic functions to be preserved across geographically distinct *M. aeruginosa* blooms. We would also expect communities to be more similar than predicted by traditional biogeographic theory, where community similarity is expected to decrease with increasing geographic distance (MacArthur 1984; Nekola and White 1999; Green and Bohannan 2006; Nemergut et al. 2013). As a first test of this prediction, we examined community composition and function of the *M. aeruginosa* bloom microbiomes from 12 lakes across four continents, addressing the specific prediction that if global *Microcystis* blooms were composed of the same taxon, the blooms would support and require a similar suite of bacterial-provided functions, thus leading to highly similar bacterial communities across *Microcystis* blooms regardless of geographic location.

Methods

Sample collection

Samples were collected from 12 lakes spanning the globe during the peak of the *Microcystis* blooms between May 2016 and July 2017 (Fig. 1). Three surface water samples were taken from each lake with a clean 1000-mL beaker or Erlenmeyer flask and set aside undisturbed for 10 min to allow the cyanobacteria to float to the surface. Concentrated *Microcystis* biomass was poured off the top of the flask or beaker through a Nitex screen (100-µm pore size) stretched between a PVC pipe and PVC coupler to collect large *Microcystis* aggregates on the screen. Each Nitex screen was rolled up using forceps and transferred into 2-mL screw cap tubes containing 1.0-mL DNA preservative (DNA/RNA Shield, Zymo Research). This process was carried out for each water sample. Tubes were stored at −20°C until shipping. They were shipped at ambient temperature and once received were stored at −20°C prior to extraction.

DNA extraction and sequencing

DNA extraction from the preserved samples was performed using Zymo Research Quick-DNA Fecal/Soil Microbe Miniprep Kits (Zymo Research) following the manufacturer's recommended protocol. The procedure involved placing the Nitex screen into the lysis tube with glass beads, followed by mixing with a Bead Beater vortex mixer (BioSpec products) for 2 min as a first step in the extraction process. Extracted DNA was collected by decanting and used as a template for amplifying bacterial 16S rRNA gene through PCR. Forward (S-D-Bact- 0341-b-S-17, 50-CCTACGGGNGGCWGCAG-30) and reverse primers (S-D-Bact-0785-a-A-21, 50-GACTACHVGGGTATCTAAT CC-30) were used for targeting the V3 and V4 regions of bacterial 16S rRNA gene (Klindworth et al. 2013). Each 50-µL PCR reaction mix contained 2-µL of template (~30ng), 2-µL of each primer (0.4-µM, final concentration), and 25-µL PCR master mix containing DreamTaq polymerase (Thermo Fisher Scientific). Thermocycler conditions were as follows: 94°C for 3 min followed by 28 cycles of 94°C for 30 s, 55°C for 60

s, and 72°C for 75 s, and a final elongation step at 72°C for 10 min. PCR products were visualized on 1% agarose gels to confirm amplification and then purified by QIAquick PCR purification kit (Qiagen) to remove primers. From each purified sample, 4-μL were added to a second PCR mixture containing barcoded primers for multiplexed Illumina sequencing. Through re-amplification for another eight cycles in the second PCR reaction, each sample received a unique "barcode" sequence as previously described (Wawrik et al. 2012). The secondary PCR products were quantified with the Qubit dsDNA BR Assay kit (Life Technologies) on a Qubit 2.0 Fluorometer. Amplicons of all samples were pooled in an equimolar amount. Pooled samples were purified using a QIAquick PCR Purification Kit (Qiagen) and requantified with the Qubit. Paired-end sequencing of the library was performed at the Oklahoma Medical Research Foundation, using the MiSeq Reagent Kit (v3) with the read length set to 2 × 300 base pairs (bp).

Shotgun metagenomics was used to profile functional potential and to recover whole genome sequences from the *Microcystis* and microbiome communities. Metagenomic sequencing was performed on two replicates per sample. A 300-bp paired-end library was constructed according to the instructions from Illumina. The libraries were sequenced on an Illumina Genome Analyzer IIx at the Oklahoma Medical Research Foundation. Eighteen metagenomes (two per lake) were multiplexed on two lanes, and a median total of ~ 48 million raw paired-end reads was obtained for each sample (range: ~ 32–96 million, due to variations in library loading).

Sequence processing and analysis

The 16S raw sequence data was processed via the QIIME pipeline (V1.9.1) (Caporaso et al. 2010), integrated with UPARSEOTU algorithm. Paired-end reads were joined using the join_paired_ends.py function according to the SeqPrep method

(https://github.com/jstjohn/SeqPrep) with a minimum overlap of 150 bp. The quality-joined fragments were demultiplexed and primer sequences removed. After using the FASTQ Quality Filter (q = 20, p = 80) to remove unqualified sequences, the remaining fragments were clustered into operational taxonomic units (OTUs) at the 97% similarity level with UPARSE OTU clustering (Edgar 2013) which generates a representative set of high-quality OTU sequences and filters out chimeric sequences via the de novo mode (using usearch 11.0.667; Edgar 2010). Taxonomic annotations were assigned to each high-quality OTU sequence by RDP's naive Bayesian rRNA Classifier (Wang et al. 2007) against the SILVA SSUv132 Reference Database (Quast et al. 2012), at the confidence threshold of 80%. To perform phylogenetic analysis, OTU sequences were also aligned against the SILVA database with PyNAST (Caporaso et al. 2009), filtered, and a phylogenetic tree constructed with FastTree (Price et al. 2010). OTUs that failed in alignment or were classified as either eukaryote, archaea, chloroplast, or mitochondria were discarded from the OTU table, as were OTUs with fewer than 100 counts summed across samples.

To compare the community structure of the *Microcystis* microbiome among lakes, we extracted non-cyanobacterial OTUs from the OTU table, and then retrieved their associated representative sequences. A phylogenetic tree with non-cyanobacterial OTU sequences was constructed, as detailed above. OTU tables without cyanobacteria were rarified to the number of reads of the sample with the fewest reads. The rarified OTU table with pooled replicates was used for all downstream diversity calculation and statistical analysis.

Following quality trimming (Trimmomatic v 0.39; Bolger et al. 2014), with reads shorter than 20 bp being discarded and removal of human genes (MetaWRAP; Uritskiy et al. 2018), the clean metagenomic data were assembled into contigs by de novo assembly of each sample

sequence using metaSPAdes (SPAdes v.3.13.0; Bankevich et al. 2012). The interactome (i.e., *Microcystis* and its microbiome) metagenomic assembled genomes (MAGs) were generated for each lake using three tools with default options: MaxBin (v.2.2.6) (Wu et al. 2015), MetaBAT (v. 2.12.1) (Kang et al. 2015), and CONCOCT (v. 1.0.0) (Alneberg et al. unpubl. preprint doi: arXiv:1312.4038v1 [q-bio.GN]), followed by integration using DAS Tool (using a threshold of ≥ 70% genome completeness) (v. 1.1.1) (Sieber et al. 2018). The complete MAGs were then divided into two groups, *Microcystis* and heterotrophic bacteria, and pooled across the lakes. Protein encoding genes of each group were annotated from the contigs with Prokka (v1.13.3) (Seemann 2014). Duplicate genes were removed with CD-HIT-EST (v 4.6) (Hahn et al. 2016). Genes were then converted into protein sequences using Prokka. The protein sequences of each group were annotated to Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologies to characterize individual gene functions using GhostKOALA (Kanehisa et al. 2016). To calculate the gene abundance in each sample, all KEGG annotated genes were first aligned with the clean reads by bowtie2 alignment software (Langmead and Salzberg 2012). The number of reads mapping to each gene was extracted using the SAMtools (v1.3.1) "idxstats" command. The abundance of each gene in all samples was calculated by get_count_table.py (https://github.com/edamame-course/Metagenome/blob/master/get_count_ table.py). To compare the functions contributed by the microbiome to those of the *Microcystis*, we constructed complete KEGG pathways (no less than one missing gene). Due to the high phylogenetic and functional similarities found across lakes (see the Results section), we pooled *Microcystis* MAGs and microbiome MAGs (Li et al. 2018), respectively, in order to generate the pathways. We uploaded the KO numbers into the online KEGG mapper to construct the complete KEGG pathways.

Due to the low depth of coverage for *Microcystis* microbiome genes, we attempted to maximize the probability of identifying functional potential in the *Microcystis* microbiomes by repeating the process outlined above for MAGs using the total metagenomic data. We separated the *Microcystis* reads from the microbiome reads by aligning the clean metagenomic data with 16 *Microcystis* genomes (10 *M. aeruginosa*, 1 *Microcystis viridis*, 1 *Microcystis panniformis*, 2 *Microcystis flos-aquae*, and 2 *Microcystis wesenbergii*) using bwa v0.7.15 (Li and Durbin 2009) and splitting the data based on this alignment. The reads were then assembled into contigs (SPAdes v3.1.1; Bankevich et al. 2012) and the contigs for each lake (*Microcystis* and the microbiome kept separate) were pooled across lakes into a single group and annotated (Prokka v1.13.3; Seemann 2014). Duplicate genes were then removed with CD-HIT-EST (v 4.6) (Hahn et al. 2016). The protein sequences of each group were annotated to KEGG orthologies, the gene abundance mapped to each lake, and the data pooled and complete KEGG pathways constructed as described above.

Raw sequence data in this study have been deposited at the National Center for Biotechnology Information website under BioProject accession number PRJNA575023.

Statistical analyses

All statistical analyses were completed in the R statistical environment v.3.5.1 (R Development Core Team 2018), except where noted otherwise. We first tested for associations between geographic distance and community dissimilarity across samples. We calculated the great circle geographic distance between sites using the "rdist.earth" function (fields v.9.6). We calculated Bray–Curtis dissimilarity using "vegdist" (vegan package v.2.5-3) (Oksanen et al. 2013). UniFrac values, weighted by abundance, were generated in QIIME. Generalized linear

models (GLM) were used to assess the relationship between geographic distance and community dissimilarity measures (using default family = Gaussian, link = identity). Deviance explained by GLMs coupled with p-values was used to assess the significance and strength of the relationship.

To examine the phylogenetic relationship among *Microcystis* microbiome bacteria within each sample (α-diversity), we calculated mean-nearest-taxon-distance (MNTD) and the nearest-taxon-index (NTI) (Webb 2000) using the "mntd" and "ses.mntd" commands in the Picante package v.1.7 in R (Kembel et al. 2010; Swenson 2014). NTD is a measure of phylogenetic distance between each OTU within a sample and its closest relative in the same sample. The mean is then calculated across all phylogenetic distances in a sample to give a value of phylogenetic relatedness. To determine whether observed phylogenetic community composition was more or less related (or structured) than predicted by chance, null models were built by randomly shuffling the taxa within each community across the tips of the phylogeny (null.model = "taxa.labels" in "ses.mntd") and recalculating MNTD 999 times (Stegen et al. 2012). The resulting NTI (which is the negative output of "ses.mntd") distribution displays the number of standard deviations that the observed MNTD is from the mean of the null MNTD values. An αNTI value less than −2 indicates taxa within the community are more distantly related than by chance (phylogenetically over dispersed), a value greater than 2 indicates taxa are more closely related than expected by chance (phylogenetically clustered), and αNTI values between 2 standard deviations of 0 indicate that the observed community is no different from random (Webb 2000; Stegen et al. 2012). A two-sample t-test was used to assess whether the mean αNTI values from across all communities was significantly different from that of the null distribution.

We used beta-mean-nearest-taxon-distance (βMNTD) and beta-nearest-taxon-index (βNTI) to quantify the phylogenetic distance among samples (see Swenson 2014 for R code).

Similar to its α-diversity analog, βMNTD is the mean of the phylogenetic distances between each OTU in a given sample to their nearest relative in the comparison sample. Null distributions were generated by randomizing OTUs across the phylogeny and recalculating βMNTD 999 times (Stegen et al. 2012). As before, βNTI is the number of standard deviations that the observed βMNTD is away from the mean of the null distribution. A value of βNTI less than −2 indicates more phylogenetic relatedness between samples than expected by chance, and a βNTI greater than 2 indicates less phylogenetic relatedness between the two communities than expected by chance (i.e., species in these communities are more phylogenetically distant from each other). A two-sample t-test was used to determine if the mean βNTI for two communities was significantly different from the null expectation.

The abundance of KEGG genes for the *Microcystis* microbiome for each lake (mapped from the pooled total metagenomic data) was used to measure the community functional dissimilarity using Bray–Curtis (pairwise comparisons between lakes). GLM were used to assess the relationship between geographic distance and community functional dissimilarity measures (using default family = Gaussian, link = identity). To explore important biogeochemical pathways that may be shared or different between the *Microcystis* and their associated bacteria, we compared the complete (no more than one gene missing) KEGG pathways between *Microcystis* and the microbiome. Due to the low depth of coverage of the microbiome, we used both the MAGs complete KEGG pathways and the total metagenome complete KEGG pathways for this analysis to maximize the probability of identifying a complete pathway.

Results

The sampled lakes spanned a 280° longitudinal and 90° latitudinal gradient ranging from 444 to 11,777 km apart and represented *M. aeruginosa* blooms from four continents (Fig. 1). After removal of Eukaryota, singletons, chloroplasts, and mitochondrial reads, 3.25 million high-quality reads were represented by 454 unique (97%similarity) bacterial OTUs. *Microcystis* (927 total OTUs) was dominant in 10 of the 12 lakes and ranged from 65% to 84% of total sequence abundance. Both Taihu and Wentowsee communities were made up of less than 50% *Microcystis* (48% and 5%, respectively) sequences suggesting the lakes were not at peak bloom phase during sampling. As such, these two lakes were removed from subsequent analyses. Castlerock was also removed from subsequent analyses due to a loss of one of the triplicate samples. All *Microcystis* OTUs with greater than 1000 total abundance were reidentified to species using NCBI BLAST function (Altschul et al. 1990) and were all identified as *M. aeruginosa*, accounting for 79–85% of the total *Microcystis* across lakes and 53–68% of all OTU abundances across lakes. This confirms that, at least at the level of 16S rRNA sequences, *M. aeruginosa* is a cosmopolitan bloom-forming species.

The 454 non-cyanobacterial OTUs were associated with 35 bacterial classes, with most OTUs identified as *Alphaproteobacteria, Bacteroidia, Gammaproteobacteria, Clostridia, Campylobacteria, Deltaproteobacteria, Negativicutes, Phycisphaerae, Gemmatimonadetes, Acidobacteriia, Ignavibacteria, SM1A07, Melainabacteria, Cytophagia, Parcubacteria*, and *Anaerolineae* (Fig. 2b). The abundance and taxonomy of the associated bacteria differed among sites (Fig. 2b) with *Alphaproteobacteria*, *Bacteriodia*, and *Gammaproteobacteria* classes contributing most sequences across sites. Bray–Curtis dissimilarity values were relatively high for all of the site comparisons indicating large differences in OTUs among sites (Fig. 3a). Weighted UniFrac, on the other hand, was generally low, suggesting that among sites, there were

many shared taxa (or a few shared taxa with high abundances) with some degree of phylogenetic relatedness (Fig. 3b). These results together suggest that there was a wide range of relative abundances of taxa across all of the sites, including many low abundance rare taxa (contributing to high Bray–Curtis values), while the highly abundant taxa were more related and shared across most sites (contributing to lower UniFrac values). Neither of the community dissimilarity metrics were significantly correlated with geographic distance (Fig. 3a,b; BC GLM deviance explained = 4.1%, $p = 0.2$; UF GLM DE = 0.97%, $p = 0.57$).

Phylogenetic α-diversity analysis using mean nearest taxon distance and null model generation revealed on average that within lakes, bacteria in the *Microcystis* microbiome were more phylogenetically related than predicted by chance (Fig. 4, +αNTI; $t = 10.13$, $p < 0.001$). Among lakes, the *Microcystis* microbiome communities also were significantly more phylogenetically related than expected by chance (Fig. 4, βNTI; $t = -12.65$, $p < 0.001$).

A total of 866 million metagenomic sequencing reads were generated from the nine lake *Microcystis*-microbiome communities. After trimming and filtering, 612 million (52.4–137.6 million per lake) clean reads were generated (Supplementary Table S1). Most of the clean metagenomic reads (55.2–76.6% per lake) belonged to *Microcystis*.

Analysis of only whole genome data revealed nine *Microcystis* genome bins (one MAG per lake) and but only 43 microbiome bacterial genome bins (3–10 MAGs per lake), indicating extremely low depth coverage for the microbiome bacteria. After removing the duplicate genes and RNA genes, 156,445 and 39,880 protein-coding genes were obtained from the microbiome and *Microcystis* genomes, respectively. Our analysis showed that 66,861 (42%) of the microbiome genes, and 14,188 (~ 35%) of *Microcystis* genes were successfully assigned to the KEGG orthology. By contrast, after removing duplicate and rRNA genes, analysis of the total

metagenomic data produced 407,658 and 54,312 protein-coding genes for the microbiome and *Microcystis*, respectively, of which 139,384 (more than twice the number compared with MAGs) and 17,817 (3629 more), respectively, were successfully mapped to the KEGG orthology. Microbiome gene diversity was similar across lakes (Fig. 5; low Bray–Curtis values) and did not differ with geographic distance between the lakes (Fig. 5; BC GLMDE = 1.97%, p = 0.41).

Shared pathways were mostly assigned to carbohydrate metabolism; amino acid, nucleotide, and fatty acid biosynthesis; and cofactor and vitamin biosynthesis (Fig. 6). The microbiome bacteria contained unique pathways for organic carbon transportation and degradation and vitamin B12 synthesis not found in *Microcystis*. Numerous pathways for organic carbon degradation (D-galacturonate, D-glucuronate, galactose, glycogen, fatty acid, purine, pyrimidine) and transportation (e.g., proline, maltose, galactose, maltose, mannose, D-xylose, fructose, rhamnose, glycerol), as well as complete pathways for degradation of aromatics (mostly anthropogenic pollutants, such as toluene, xylene, benzene, phthalate) were also identified in the microbiome. Numerous important anaerobic bacterial pathways, including nitrogen fixation, denitrification, and dissimilatory sulfate reduction were also detected in the *Microcystis* microbiome, suggesting the anaerobic bacteria play an important role in nutrient cycling during *Microcystis* blooms in these lakes. With the exception of some genes related to carbon cycling (purine, pyrimidine, salicylate, and catechol degradation), as well as bacterial pathways involved in methane metabolism (methane oxidation and formaldehyde assimilation), the majority of potential biochemical function appeared to be associated with the 43 microbiome MAGs (Fig. 6 and Table S2).

Discussion

Genome reduction leads to a loss of function (Giovannoni et al. 2014), necessitating interactions with community members capable of carrying out those functions (Morris et al. 2012; Garcia et al. 2015). With small genomes compared to other algae (e.g., 4.2 vs. 34.5 Mbp, Armbrust 2004; Gregory et al. 2007), *Microcystis* spp. are potentially missing some key metabolic functions (Steffen et al. 2012) and might be reliant on community members to fill in the metabolic gaps. Similar potential mutualist interactions have recently been mapped between *Microcystis* and their microbiome of associated bacteria (Xie et al. 2016; Li et al. 2018), corroborating the original idea of a "phycosphere" of functional interaction within the *Microcystis* aggregates (Bell and Mitchell 1972; Paerl and Kellar 1978, 1979; Paerl and Millie 1996). Moreover, bacteria have been shown to complement algae in marine ecosystems by excreting large amounts of exometabolites including growth factors and biosynthetic precursors, as well as processing toxic metabolites (Morris et al. 2011; Pérez et al. 2016; Lee et al. 2017; Wienhausen et al. 2017). Given the predominance of such data across aquatic system, we hypothesized *M. aeruginosa* blooms have a microbiome of certain species of associated bacteria or perhaps metabolic functions that will be preserved across geographically distinct *Microcystis* blooms.

M. aeruginosa was the most abundant *Microcystis* species across all lakes and this confirms that, at least at the level of 16S rRNA sequences, *M. aeruginosa* is a cosmopolitan bloom-forming species. We found remarkable phylogenetic relatedness among associated bacteria, and similar function between sites, despite those bacteria being taxonomically distinct at the 16S rRNA level. We found no relationship between community composition dissimilarity and geographic distance (Fig. 3a, b), indicating no distance-decay relationship as would be expected for dispersal-limited species (Nekola and White 1999; Green and Bohannan 2006;

Nemergut et al. 2013). We also conclude that the functional potential of microbial communities is more highly conserved than their taxonomic composition (Figs. 3a, 5). Similarly, Steffen et al. (2012) found that the cyanobacterial bloom-associated bacterial communities across three lakes (Erie, Taihu, St.Marys) differed taxonomically, while being functionally similar. Thus, while OTU identity was variable across *Microcystis* microbiome bacteria, functional potential appears to have been quite similar providing evidence that under bloom conditions, *M. aeruginosa* is accompanied by a common suite of bacterial functionality, potentially forming an interactome.

We found that *Microcystis* microbiome bacteria have functional potential not found in *Microcystis*, and functional similarity was preserved globally. Pathways involving photosynthesis (excluding the anoxygenic photosystem II pathway) and carbon fixation were only detected in *Microcystis* (Fig. 6). The microbiome bacteria could potentially be contributing to carbon recycling within the aggregates as many of the pathways found only in the bacteria are related to carbohydrate breakdown (e.g., d-galacturonate, galactose, and glycogen degradation) and transport (e.g., numerous transport systems). The microbiome bacteria are likely tightly associated with the carbon sources found in the *Microcystis* aggregates. Bacteria associated with phytoplankton blooms have been found to utilize the carbon source glycolate from phytoplankton (Lau et al. 2007; Paver and Kent 2010) and contain the glycolate oxidation pathway. Unfortunately, the KEGG database does not contain this gene or related pathways thus we were unable to detect it in our samples. Every metagenomic function database has its shortcomings, and for future studies we would recommend using multiple databases to cover as many genes and pathways as possible. For example, using the COG database (clusters of orthologous groups) to search for specific genes, we in fact found that the glcD gene, which is involved in the glycolate oxidation pathway, as well as the microcystin production genes (mcyA-

I) were present in the microbiome and the *Microcystis*, respectively, in all of our lakes. In addition, we found that the microbiome potentially contributes methane metabolism pathways (methane oxidation and formaldehyde assimilation) which are used to convert methane into a useable carbon form. Recently, cyanobacteria have been suggested to produce methane during blooms (M. Bižić et al. unpubl preprint, https://doi.org/10.1101/398958), although methane is typically produced by methanogenic archaea. While we did not analyze the archaea in our samples, other studies have found methanogenic archaea can be closely associated with cyanobacteria blooms (Batista et al. 2018), thus we would expect some methane production within aggregates.

In addition, the ethylmalonyl pathway was identified in the microbiome bacterial genomes. The ethylmalonyl pathway is a new acetate assimilation strategy in *Rhodobacter* sphaeroides, an anoxygenic phototrophic organism that lacks the key enzyme of the glyoxylate cycle, isocitrate lyase (Erb et al. 2007). Indeed, genes associated with anoxygenic photosystem II were identified in six microbiome MAGs (data not shown). Sulfur is a byproduct of anoxygenic photosynthesis and we also found evidence for thiosulfate oxidation and dissimilatory sulfate reduction pathways present in the microbiome bacteria. These results corroborate Li et al. (2018), who suggested that bacteria associated with *Microcystis* (termed epibionts by Li et al.) were essential for maintaining the redox balance and cycling different forms of sulfur within *Microcystis* aggregates. In addition, the vitamin B_{12} biosynthesis pathway, a necessary vitamin that *Microcystis* cannot produce, was only detected in the microbiome bacteria. Croft et al. (2005, 2006) proposed that most phytoplankton were likely auxotrophic for vitamin B_{12} and other essential vitamins. Indeed, previous studies have also suggested that vitamin B_{12} was

provided to *Microcystis* by the associated bacteria and, moreover, that this relationship was mutually beneficial for both groups (Xie et al. 2016; Li et al. 2018).

Most *Microcystis*-blooming lakes are typically recipients of urban and agricultural runoff. Correspondingly, we found the *Microcystis* microbiome contained the degradation pathways for many potentially harmful aromatic pollutants (e.g., benzene, benzoate, phthalate, etc.). *Microcystis* could be benefiting from pollutant degradation as many of these aromatics have been shown to inhibit phytoplankton growth (Häder and Gao 2015). Xie et al. (2016) also found a group of aggregate-associated bacteria that contributed the whole benzoate degradation pathway to the community, again pointing toward mutualism between the *Microcystis* and the *Microcystis* microbiome. Together, these results support our hypothesis of a coevolved interactome. We also hypothesize that many of these microbiome functions (Fig. 6) are needed for growth and dominance by *M. aeruginosa* during bloom conditions.

Loss of necessary metabolic functions is more common in bacterial communities than previously thought (Hottes et al. 2013). Morris et al. (2012) found that the marine cyanobacterium, *Prochlorococcus*, has lost many oxidative-stress genes and instead relies on other microbes for removal of hydrogen peroxide from the microenvironment. As such, *Prochlorococcus* may benefit from hydrogen peroxide removal via the actions of "helper" microbes in the form of a "leaky" public good (sensu Morris et al. 2012), and the smaller genome of *Prochlorococcus* affords it a selective advantage. A recent genomic study has shown that *M. aeruginosa* also has a reduced genome with remarkable redundancy consisting of a set of ~ 2400 core genes and a large, variable pangenome, an additional set of genes unique to different *M. aeruginosa* strains likely acquired through horizontal gene transfer (Humbert et al. 2013). We speculate that the variability within the *M. aeruginosa* pangenome could be due to differences in

the need for specific functions across different environments. Horizontal gene transfer and the Black Queen Hypothesis provide two mechanisms for coping with changing environmental needs. “Leaky” functions may be lost from *Microcystis* when they are costly and a public good is available, as with the case of oxidative-stress genes in *Prochlorococcus*. However, when that public good becomes less predictable, those functions may be recouped through horizontal gene transfer if the new conditions provide a selective advantage to carriers for regaining the function. Steffen et al. (2012) corroborated this hypothesis as they found microbiome functional potential remained static between two *Microcystis* blooms. However, in one of the lakes (Taihu), *Microcystis* was reliant on Proteobacteria for nitrogen assimilation and metabolism, while in the other lakes (Erie and St. Marys), *Microcystis* carried out those functions. In the present study, we also found that the microbiome bacteria were potentially the sole contributors of nitrogen fixation, denitrification, and the urea cycle pathways (Fig. 6). *Microcystis* cannot fix nitrogen so this may indicate that *Microcystis* is relying on public goods across multiple blooms. This interactome of *M. aeruginosa* and its microbiome could be an ideal system to test for “Black Queen” functions.

Alternatively, the similarities we see in the microbiome bacterial community could be due in part to the *Microcystis* bloom creating specific habitats that are selective of certain types of heterotrophic bacteria. As previously mentioned, *Microcystis* blooms can change local environments, creating large amounts of particulate organic matter that have been shown to be an important nutrient source for shaping the bacterial community (e.g., Fogg and Harold 1952; Bell and Mitchell 1972; Paerl and Gallucci 1985). We saw through βNTI analysis that *Microcystis* microbiome communities were significantly phylogenetically related (Fig. 4), with the negative value of βNTI indicating a common environmental filter across the lakes—*Microcystis* (sensu

Dini-Andreote et al. 2015). Similarly, multiple studies (Yang et al. 2017; Shi et al. 2018; Xu et al. 2018) have shown that the *Microcystis*-associated (i.e., particle attached, which is equivalent to our use on *Microcystis* microbiome) bacterial community composition appeared to be heavily structured by the bloom compared to free-living bacteria. Over the course of a *Microcystis* bloom, Parveen et al. (2013) found that *Microcystis* bloom aggregates provided habitat for a bacterial community distinct from the free-living bacteria. In addition, specific bacteria have also been found utilize parts of bloom aggregates. For example, bacteria in the genus, *Sphingomonas*, actively break down toxins while associated with *Microcystis* blooms (Dziallas and Grossart 2011). This bloom-as-habitat hypothesis could also explain the differences in attached and free-living bacteria described in previous studies (Yang et al. 2017; Shi et al. 2018; Xu et al. 2018). This hypothesis does not negate that *Microcystis* could be exchanging or receiving functions from these aggregate-associated bacteria. Additional metagenomic and multi-year bloom studies are needed to further parse these relationships.

The phylogenetic and functional similarities compared with the taxonomic dissimilarities we observed provide support to a growing body of evidence suggesting that community composition comparisons should be based on functional genes rather than strictly taxonomy (OTUs) (Burke et al. 2011; Oh et al. 2011). While clear similarities are shown using 16S rRNA genes, this method is taxonomically conservative and multiple different bacterial species could be represented by a single OTU. 16S rRNA genes represent a very small part of the genome, and evidence is growing which suggests that 16S rRNA is insufficient for distinguishing freshwater bacteria. For example, Polynucleobacter species, with 16S rRNA similarities $\geq$ 99%, have been shown to be distinct species based on whole genome comparisons and ecological isolation (Hahn et al. 2016). In *Microcystis*, a threshold of 98–99% similarity is suggested to be sufficient in

distinguishing some species (Harke et al. 2016). In our study, all *Microcystis* OTUs with greater than 1000 sequence abundance were assigned to *M. aeruginosa* at the species level, thus we are confident in this taxonomy, but also recognize the need for more detailed taxonomic analysis to further investigate the diversity of *Microcystis* within a bloom beyond the 97% similarity threshold for 16S rRNA genes.

In the current study, we have used 16S rRNA taxonomy paired with metagenomic functional analysis to describe the community composition and potential function of the associated bacterial community in nine geographically distinct *Microcystis* blooms. The multiple samples across continents have allowed us to confirm that *M. aeruginosa* is a cosmopolitan bloom former. The phylogenetic and functional similarity of associated bacteria across sites and the sets of complementary pathways found between the *Microcystis* and associated bacteria support the presence of a synergistic interactome. These results highlight the need for deeper investigation into *Microcystis* taxonomic identity and both *Microcystis* and microbiome functional capabilities at different times before, during, and after a bloom to fully elucidate the interactome relationship between *M. aeruginosa* and its microbiome. We conclude that the coevolution within an interactome, that is, between *M. aeruginosa* and its microbiome, could help explain the global distribution of, and dominance by, *M. aeruginosa* in diverse freshwater ecosystems.

Acknowledgments

We thank the members of the University of Oklahoma Plankton Ecology and Limnology Laboratory for helpful discussions, lab and statistical assistance, and for commenting on earlier versions of this manuscript. In particular, we thank Jessica Beyer for invaluable statistical advice and critical editorial feedback and Stephen C. Cook for making Fig. 1. Funding was provided by the University of Oklahoma Office of the Vice President for Research to L.R.K. and K.D.H., by the National Science Foundation (DEB 1831061) to K.D.H., L.R.K., H.W.P., M.M.S., and A.E.W., NSF OCE 1840715 and NIH-1P01ES028939-01 to H.W.P, NSF DEB 1831106 to M.M.S., and through an NSF Graduate Research Fellowship to K.V.C. This work will be submitted by K.V.C. as partial fulfillment of the requirements for a Ph.D. degree from the University of Oklahoma.

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403–410. doi:10.1016/S0022-2836(05)80360-2

Armbrust, E. V. 2004. The genome of the diatom Thalassiosira pseudonana: Ecology, evolution, and metabolism. Science 306: 79–86. doi:10.1126/science.1101156

Bankevich, A., and others. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19: 455–477. doi:10.1089/cmb.2012.0021

Batista, A., J. Woodhouse, H.-P. Grossart, and A. Giani. 2018. Methanogenic archaea associated to *Microcystis* sp. in field samples and in culture. Hydrobiologia 831: 1–10. doi:10.1007/s10750-018-3655-3

Bell, W., and R. Mitchell. 1972. Chemotactic and growth responses of marine bacteria to algal extracellular products. Bio. Bull. 143: 265–277. doi:10.2307/1540052

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120. doi:10.1093/bioinformatics/btu170

Bourne, D. G., P. G. Dennis, S. Uthicke, R. M. Soo, G. W. Tyson, and N. Webster. 2013. Coral reef invertebrate microbiomes correlate with the presence of photosymbionts. ISME J. 7: 1452–1458. doi:https://doi.org/10.1038/ismej.2012.172

Burke, C., P. Steinberg, D. Rusch, S. Kjelleberg, and T. Thomas. 2011. Bacterial community assembly based on functional genes rather than species. PNAS 108: 14288–14293. doi:10.1073/pnas.1101591108

Cai, H., H. Jiang, L. R. Krumholz, and Z. Yang. 2014. Bacterial community composition of size-fractioned aggregates within the phycosphere of cyanobacterial blooms in a eutrophic freshwater lake. PLoS One 9: e102879. doi:10.1371/journal.pone.0102879

Caporaso, J. G., K. Bittinger, F. D. Bushman, T. Z. DeSantis, G. L. Andersen, and R. Knight. 2009. PyNAST: A flexible tool for aligning sequences to a template alignment. Bioinformatics 26: 266–267. doi:10.1093/bioinformatics/btp636

Caporaso, J. G., and others. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7: 335–336. doi:10.1038/nmeth.f.303

Carey, C. C., B. W. Ibelings, E. P. Hoffmann, D. P. Hamilton, and J. D. Brookes. 2012. Eco-physiological adaptations that favour freshwater cyanobacteria in a changing climate. Water Res. 46: 1394–1407. doi:10.1016/j.watres.2011.12.016

Carmichael, W. W. 2001. Health effects of toxin-producing cyanobacteria: “The CyanoHABs”. Hum. Ecol. Risk. Assess. 7: 1393–1407. doi:10.1080/20018091095087

Conley, D. J., H. W. Paerl, R. W. Howarth, D. F. Boesch, S. P. Seitzinger, K. E. Havens, C. Lancelot, and G. E. Likens. 2009. Controlling eutrophication: Nitrogen and phosphorus. Science 323: 1014–1015. doi:10.1126/science.1167755

Croft, M. T., A. D. Lawrence, E. Raux-Deery, M. J. Warren, and A. G. Smith. 2005. Algae acquire vitamin B12 through a symbiotic relationship with bacteria. Nature 438: 90–93. doi:10.1038/nature04056

Croft, M. T., M. J. Warren, and A. G. Smith. 2006. Algae need their vitamins. Eukaryot. Cell 5:1175–1183. doi:10.1128/ EC.00097-06

Dini-Andreote, F., J.C. Stegen, J. D. vanElsas, and J. F. Salles. 2015. Disentangling mechanisms that mediate the balance between stochastic and deterministic processes in microbial succession. PNAS 112: E1326–E1332. doi:10.1073/pnas.1414261112

Dziallas, C., and H. Grossart. 2011. Temperature and biotic factors influence bacterial communities associated with the cyanobacterium *Microcystis* sp. Environ. Microbiol.. 13: 1632–1641. dio:10.1111/j.1462-2920.2011.02479.x

Dziallas, C., and H.-P. Grossart. 2012. Microbial interactions with the cyanobacterium *Microcystis aeruginosa* and their dependence on temperature. Mar. Biol. 159: 2389–2398. doi:10.1007/s00227-012-1927-4

Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460–2461. doi:10. 1093/bioinformatics/btq461

Edgar, R. C. 2013. UPARSE: Highly accurate OTU sequences from microbial amplicon reads. Nat. Methods 10: 996–998. doi:10.1038/nmeth.2604

Eiler, A., and S. Bertilsson. 2004. Composition of freshwater bacterial communities associated with cyanobacterial blooms in four Swedish lakes. Environ. Mircobiol. 6: 1228–1243. doi:10.1111/j.1462-2920.2004.00657.x

Erb, T. J., I. A. Berg, V. Brecht, M. Muller, G. Fuchs, and B. E. Alber. 2007. Synthesis of C5-dicarboxylic acids from C2-units involving crotonyl-CoA carboxylase/reductase: The ethylmalonyl-CoA pathway. PNAS 104: 10631–10636. doi:10.1073/pnas.0702791104

Fierer, N. 2017. Embracing the unknown: Disentangling the complexities of the soil microbiome. Nat. Rev. Microbiol. 15: 579–590. doi:10.1038/nrmicro.2017.87

Fogg, G. E., and P. W. Harold. 1952. The production of extracellular nitrogenous substances by a blue-green alga. Proc. R. Soc. B 139: 372–397. doi:10.1098/rspb.1952.0019

Garcia, S. L., M. Buck, K. D. McMahon, H.-P. Grossart, A. Eiler, and F. Warnecke. 2015. Auxotrophy and intrapopulation complementary in the “interactome” of a cultivated freshwater model community. Mol. Ecol. 24: 4449–4459. doi:10.1111/mec.13319

Giovannoni, S. J., J. Cameron Thrash, and B. Temperton. 2014. Implications of streamlining theory for microbial ecology. ISME J. 8: 1553–1565. doi:10.1038/ismej.2014.60

Green, J., and B. J. M. Bohannan. 2006. Spatial scaling of microbial biodiversity. Trends Ecol. Evol. 21: 501–507. doi:10.1016/j.tree.2006.06.012

Gregory, T. R., and others. 2007. Eukaryotic genome size databases. Nucleic Acids Res. 35: D332–D338. doi:10.1093/nar/gkl828

Häder, D.-P., and K., Gao. 2015. Interactions of anthropogenic stress factors on marine phytoplankton. Front. Environ. Sci 3:14. doi:10.3389/fenvs.2015.00014

Hahn, M. W., J. Jezberová, U. Koll, T. Saueressig-Beck, and J. Schmidt. 2016. Complete ecological isolation and cryptic diversity in Polynucleobacter bacteria not resolved by 16S rRNA gene sequences. ISME J. 10: 1642–1655. doi:10.1038/ismej.2015.237

Harke, M. J., M. M. Steffen, C. J. Gobler, T. G. Otten, S. W. Wilhelm, S. A. Wood, and H. W. Paerl. 2016. A review of the global ecology, genomics, and biogeography of the toxic cyanobacterium, *Microcystis* spp. Harmful Algae 54:4–20. doi:10.1016/j.hal.2015.12.007

Havens, K. E. 2008. Cyanobacteria blooms: Effects on aquatic ecosystems, p. 733–747. In H. K. Hudnell [ed.], Cyanobacterial harmful algal blooms: State of the science and research needs. New York: Springer. doi:10.1007/978-0-387-75865-7_33

Herdman, M., M. Janvier, R. Rippka, and R. Y. Stanier. 1979. Genome size of cyanobacteria. J. Gen. Microbiol. 111:73–85. doi:10.1099/00221287-111-1-73

Hottes, A. K., P. L. Freddolino, A. Khare, Z. N. Donnell, J. C. Liu, and S. Tavazoie. 2013. Bacterial adaptation through loss of function. PLoS Genet. 9: e1003617. doi:10.1371/journal.pgen.1003617

Huisman, J., G. A. Codd, H. W. Paerl, B. W. Ibelings, J. M. H. Verspagen, and P. M. Visser. 2018. Cyanobacterial blooms. Nat. Rev. Microbiol. 16: 471–483. doi:10.1038/s41579-018-0040-1

Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486: 207–214. doi:10.1038/nature11234

Humbert, J.-F., and others. 2013. A tribute to disorder in the genome of the bloom-forming freshwater cyanobacterium *Microcystis aeruginosa*. PLoS One 8: e70747. doi:10.1371/journal.pone.0070747

Kanehisa, M., Y. Sato, and K. Morishima. 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J. Mol. Biol. 428: 726–731. doi:10.1016/j.jmb.2015.11.006

Kang, D. D., J. Froula, R. Egan, and Z. Wang. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3:e1165. doi:10.7717/peerj.1165

Kembel, S. W., P. D. Cowan, M. R. Helmus, W. K. Cornwell, H. Morlon, D. D. Ackerly, S. P. Blomberg, and C. O. Webb. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26: 1463–1464. doi:10.1093/bioinformatics/btq166

Klindworth, A., E. Pruesse, T. Schweer, J. Peplies, C. Quast, M. Horn, and F. O. Glöckner. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next generation sequencing-based diversity studies. Nucleic Acids Res. 41: e1–e1. doi:10.1093/nar/gks808

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357–359. doi:10.1038/nmeth.1923

Lau, W. W., R. G. Keil, and E. V. Armbrust. 2007. Succession and diel transcriptional response of the glycolate-utilizing component of the bacterial community during a spring phytoplankton bloom. Appl. Environ. Microbiol. 73:2440–2450. doi:10.1128/AEM.01965-06

Lee, M. D., N. G. Walworth, E. L. McParland, F.-X. Fu, T. J. Mincer, N. M. Levine, D. A. Hutchins, and E. A. Webb. 2017. The Trichodesmium consortium: Conserved heterotrophic co-occurrence and genomic signatures of potential interactions. ISME J. 11: 1813–1824. doi:10.1038/ismej.2017.49

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760. doi:10.1093/bioinformatics/btp324

Li, H., P. Xing, M. Chen, Y. Bian, and Q. L. Wu. 2011. Short-term bacterial community composition dynamics in response to accumulation and breakdown of *Microcystis* blooms. Water Res. 45: 1702–1710. doi:10.1016/j.watres.2010.11.011

Li, Q., and others. 2018. A large-scale comparative metagenomic study reveals the functional interactions in six bloom-forming *Microcystis*-epibiont communities. Front. Microbiol. 9: 746. doi:10.3389/fmicb.2018.00746

MacArthur, R. H. 1984. Geographical ecology: Patterns in the distribution of species. Princeton Univ. Press. Mantzouki, E., and others. 2018. Temperature effects explain continental scale distribution of cyanobacterial toxins. Toxins 10: 156. doi:10.3390/toxins10040156

Morris, J. J., Z. I. Johnson, M. J. Szul, M. Keller, and E. R. Zinser. 2011. Dependence of the cyanobacterium *Prochlorococcus* on hydrogen peroxide scavenging microbes for growth at the ocean's surface. PLoS One 6: e16805. doi:10.1371/journal.pone.0016805

Morris, J. J., R. E. Lenski, and E. R. Zinser. 2012. The Black Queen Hypothesis: Evolution of dependencies through adaptive gene loss. MBio 3: e00036–e00012. doi:10.1128/mBio.00036-12

Nekola, J. C., and P. S. White. 1999. The distance decay of similarity in biogeography and ecology. J. Biogeogr. 26:867–878. doi:10.1046/j.1365-2699.1999.00305.x

Nemergut, D. R., and others. 2013. Patterns and processes of microbial community assembly. Microbiol. Mol. Biol. Rev. 77: 342–356. doi:10.1128/MMBR.00051-12

Oh, S., A. Caro-Quintero, D. Tsementzi, N. DeLeon-Rodriguez, C. Luo, R. Poretsky, and K. T. Konstantinidis. 2011. Metagenomic insights into the evolution, function, and complexity

of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. Appl. Environ. Microbiol. 77: 6000–6011. doi:10.1128/AEM.00107-11

Oksanen, J., and others. 2013. Vegan: Community ecology package. R Package version 2.5-3. Available from http://CRAN.R-project.org/package=vegan

O'Neil, J. M., T. W. Davis, M. A. Burford, and C. J. Gobler. 2012. The rise of harmful cyanobacteria blooms: The potential roles of eutrophication and climate change. Harmful Algae 14: 313–334. doi:10.1016/j.hal.2011.10.027

Paerl, H. W., and K. K. Gallucci. 1985. Role of chemotaxis in establishing a specific nitrogen-fixing cyanobacterial bacterial association. Science 227: 647–649. doi:10.1126/science.227.4687.647

Paerl, H. W., and P. E. Kellar. 1978. Significance of bacterial- Anabaena (cyanophyceae) associations with respect to N2 fixation in freshwater. J. Phycol. 14: 254–260. doi:10.1111/j.1529-8817.1978.tb00295

Paerl, H. W., and P. E. Kellar. 1979. Nitrogen-fixing Anabaena: Physiological adaptations instrumental in maintaining surface blooms. Science 204: 620–622. doi:10.1126/science.204.4393.620

Paerl, H. W., and D. F. Millie. 1996. Physiological ecology of toxic aquatic cyanobacteria. Phycologia 35: 160–167. doi:10.2216/i0031-8884-35-6S-160.1

Paerl, H. W., and T. G. Otten. 2013. Harmful cyanobacterial blooms: Causes, consequences, and controls. Microb. Ecol. 65: 995–1010. doi:10.1007/s00248-012-0159-y

Paerl, H. W., and T. G. Otten. 2016. Duelling "CyanoHABs": Unravelling the environmental drivers controlling dominance and succession among diazotrophic and non-N2-fixing harmful cyanobacteria. Environ. Mircobiol. 18:316–324. doi:10.1111/1462-2920.13035

Paerl, H. W., and V. J. Paul. 2012. Climate change: Links to global expansion of harmful cyanobacteria. Water Res. 46:1349–1363. doi:10.1016/j.watres.2011.08.002

Paerl, H. W., and others. 2016. It takes two to tango: When and where dual nutrient (N & P) reductions are needed to protect lakes and downstream ecosystems. Environ. Sci. Technol. 50: 10805–10813. doi:10.1021/acs.est.6b02575

Paerl, H. W., H. Xu, M. J. McCarthy, G. Zhu, B. Qin, Y. Li, and W. S. Gardner. 2011. Controlling harmful cyanobacterial blooms in a hyper-eutrophic lake (Lake Taihu, China): The need for a dual nutrient (N & P) management strategy. Water Res. 45: 1973–1983. doi:10.1016/j.watres.2010.09.018

Pande, S., and C. Kost. 2017. Bacterial unculturability and the formation of intercellular metabolic networks. Trends Microbiol. 25: 349–361. doi:10.1016/j.tim.2017.02.015

Parveen, B., V. Ravet, C. Djediat, I. Mary, C. Quiblier, D. Debroas, and J.-F. Humbert. 2013. Bacterial communities associated with *Microcystis* colonies differ from free-living communities living in the same ecosystem: Bacterial diversity inside *Microcystis* colonies. Environ. Mircobiol. Rep. 5:716–724. doi:10.1111/1758-2229.12071

Paver, S. F., and A. D. Kent. 2010. Temporal patterns in glycolate-utilizing bacterial community composition correlate with phytoplankton population dynamics in humic lakes. Microb. Ecol. 60: 406–418. doi:10.1007/s00248-010-9722-6

Pérez, A. A., D. A. Rodionov, and D. A. Bryant. 2016. Identification and regulation of genes for cobalamin transport in the cyanobacterium Synechococcus sp. strain PCC 7002. J. Bacteriol. 198:2753–2761. doi:10.1128/JB.00476-16

Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. FastTree 2– approximately maximum-likelihood trees for large alignments. PLoS One 5: e9490. doi:10.1371/journal.pone.0009490

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner. 2012. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 41:D590–D596. doi:10.1093/nar/gks1219, D1

R Development Core Team. 2018. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Salomon, P. S., S. Janson, and E. Granéli. 2003. Molecular identification of bacteria associated with filaments of Nodularia spumigena and their effect on the cyanobacterial growth. Harmful Algae 2: 261–272. doi:10.1016/S1568-9883(03)00045-3

Schindler, D. W. 2012. The dilemma of controlling cultural eutrophication of lakes. Proc. R. Soc. B 279: 4322–4333. doi:10.1098/rspb.2012.1032

Schindler, D. W., S. R. Carpenter, S. C. Chapra, R. E. Hecky, and D. M. Orihel. 2016. Reducing phosphorus to curb lake eutrophication is a success. Environ. Sci. Technol. 50:8923–8929. doi:10.1021/acs.est.6b02204

Seemann, T. 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics 30: 2068–2069. doi:10.1093/bioinformatics/btu153

Shi, L., and others. 2018. Large buoyant particles dominated by cyanobacterial colonies harbor distinct bacterial communities from small suspended particles and free-living bacteria in the water column. MicrobiologyOpen 7:e00608. doi:10.1002/mbo3.608

Sieber, C. M., A. J. Probst, A. Sharrar, B. C. Thomas, M. Hess, S. G. Tringe, and J. F. Banfield. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat. Microbiol. 3: 836–843. doi:10.1038/s41564-018-0171-1

Steffen, M. M., Z. Li, T. C. Effler, L. J. Hauser, G. L. Boyer, and S. W. Wilhelm. 2012. Comparative metagenomics of toxic freshwater cyanobacteria bloom communities on two continents. PLoS One 7: e44002. doi:10.1371/journal.pone.0044002

Stegen, J. C., X. Lin, A. E. Konopka, and J. K. Fredrickson. 2012. Stochastic and deterministic assembly processes in subsurface microbial communities. ISME J. 6: 1653–1664. doi:10.1038/ismej.2012.22

Steiner, K., S. A. Wood, J. Puddick, I. Hawes, D. R. Dietrich, and D. P. Hamilton. 2017. A comparison of bacterial community structure, activity and microcystins associated with formation and breakdown of a cyanobacterial scum. Aquat. Microb. Ecol. 80: 243–256. doi:10.3354/ame01852

Swan, B. K., and others. 2013. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. PNAS 110: 11463–11468. doi:10.1073/pnas.1304246110

Swenson, N. G. 2014. Functional and phylogenetic ecology in R. Springer. doi:10.1007/978-1-4614-9542-0

Uritskiy, G. V., J. DiRuggiero, and J. Taylor. 2018. MetaWRAP—A flexible pipeline for genome-resolved metagenomic data analysis. Microbiome 6: 158. doi:10.1186/s40168-018-0541-1

Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73: 5261–5267. doi:10.1128/AEM.00062-07

Wawrik, B., and others. 2012. Field and laboratory studies on the bioconversion of coal to methane in the San Juan Basin. FEMS Microbiol. Ecol. 81: 26–42. doi:10.1111/j.1574-6941.2011.01272

Webb, C. O. 2000. Exploring the phylogenetic structure of ecological communities: An example for rain Forest trees. Am. Nat. 156: 145–155. doi:10.1086/303378

Wienhausen, G., B. E. Noriega-Ortega, J. Niggemann, T. Dittmar, and M. Simon. 2017. The exometabolome of two model strains of the Roseobacter group: A marketplace of microbial metabolites. Front. Microbiol. 8: 1985. doi:10.3389/fmicb.2017.01985

Wu, Y.-W., B. A. Simmons, and S. W. Singer. 2015. MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32:605–607. doi:10.1093/bioinformatics/btv638

Xie, M., M. Ren, C. Yang, H. Yi, Z. Li, T. Li, and J. Zhao. 2016. Metagenomic analysis reveals symbiotic relationship among bacteria in *Microcystis*-dominated community. Front. Microbiol. 7: 56. doi:10.3389/fmicb.2016.00056

Xu, H., and others. 2018. Contrasting network features between free-living and particle-attached bacterial communities in Taihu Lake. Microb. Ecol. 76: 303–313. doi:10.1007/s00248-017-1131-7

Yang, C., and others. 2017. Distinct network interactions in particle-associated and free-living bacterial communities during a *Microcystis aeruginosa* bloom in a Plateau Lake. Front. Microbiol. 8: 1202. doi:10.3389/fmicb.2017.01202

FIGURES

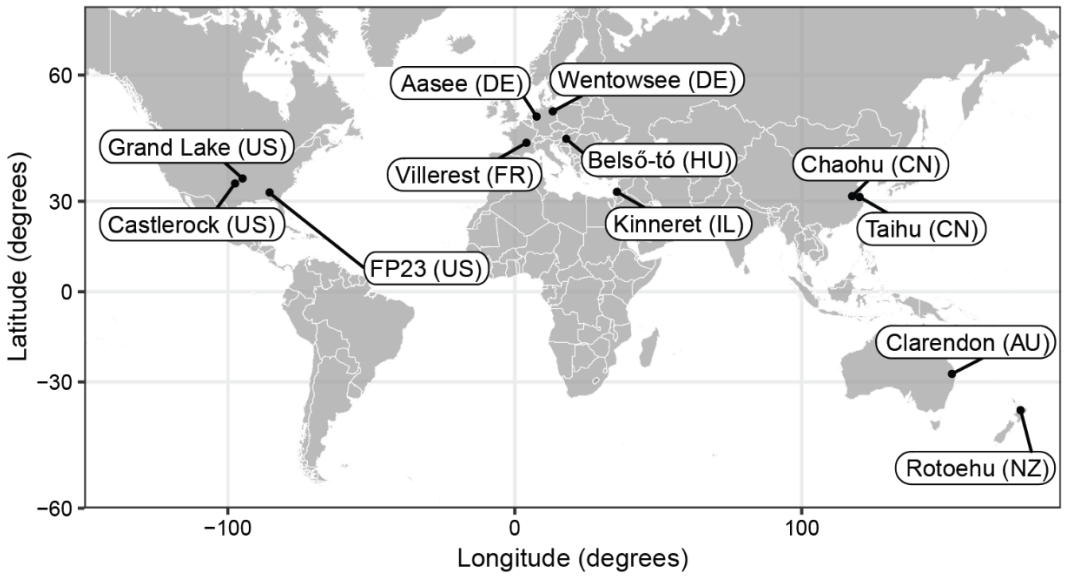


Figure 1. Location of the 12 lakes across the globe. These samples represent a 280° longitudinal and 90° latitudinal gradient.

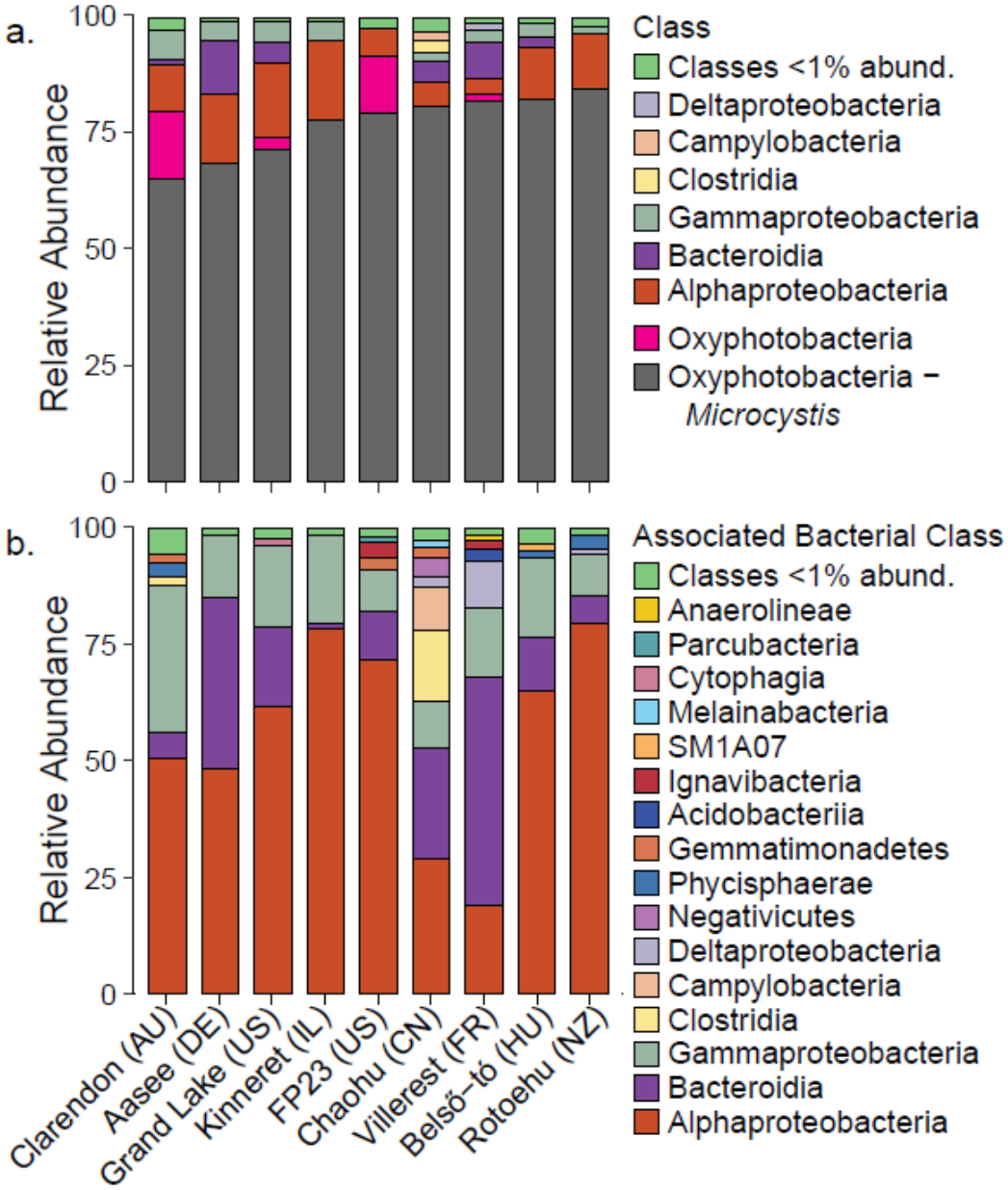


Figure 2. (A) Relative abundance of Bacteria classes in the nine lakes. The lakes are arranged in order from left to right of increasing percent *Microcystis* in the community. Classes less than 1% of the total relative abundance were grouped together as a single group denoted "<1% abund". *Oxyphotobacteria* (cyanobacteria) were split into two groups: *Microcystis* only in one and all other cyanobacteria in the second. (B) Relative abundance of non-*Microcystis* (i.e., microbiome) bacterial classes.

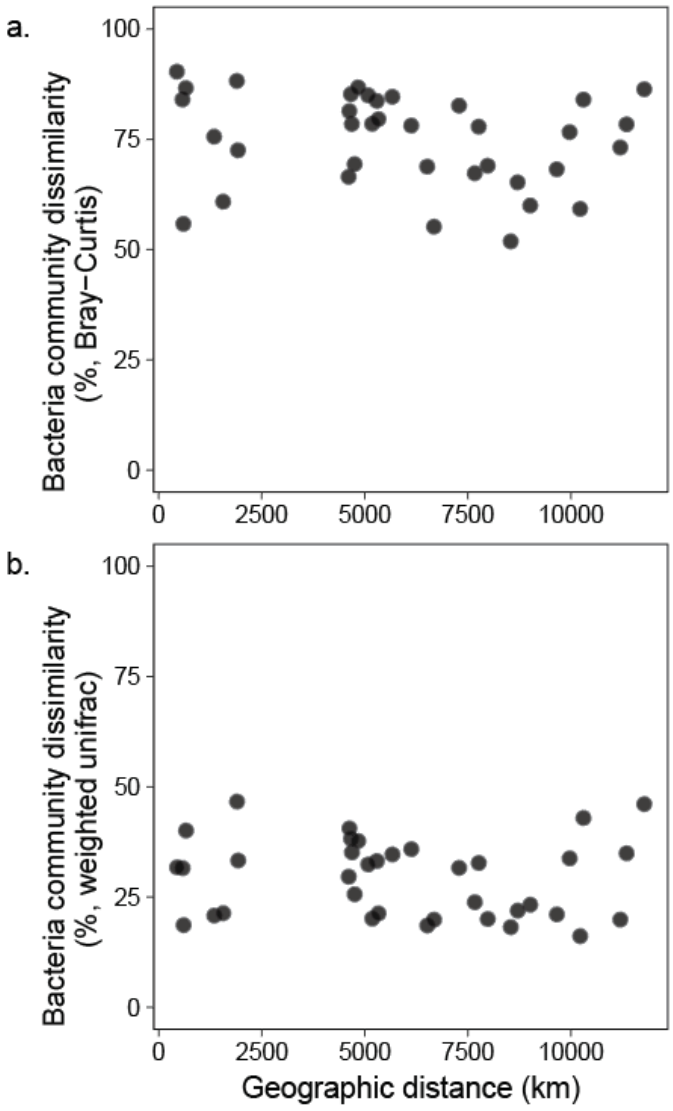


Figure 3. Scatter plots of community dissimilarity in the microbiome as related to geographic distance. (A) The non-significant (GLM Deviance explained = 3.7%, p = 0.2) relationship between taxonomic Bray-Curtis dissimilarity and geographic distance where the higher Bray-Curtis values indicate fewer species in common between sites. (B) Abundance weighted UniFrac did not scale significantly by geographic distance (GLM DE = 0.82%, p = 0.55). Here higher values of UniFrac indicate there is little overlap in species between communities whereas lower values indicate the communities are more similar.

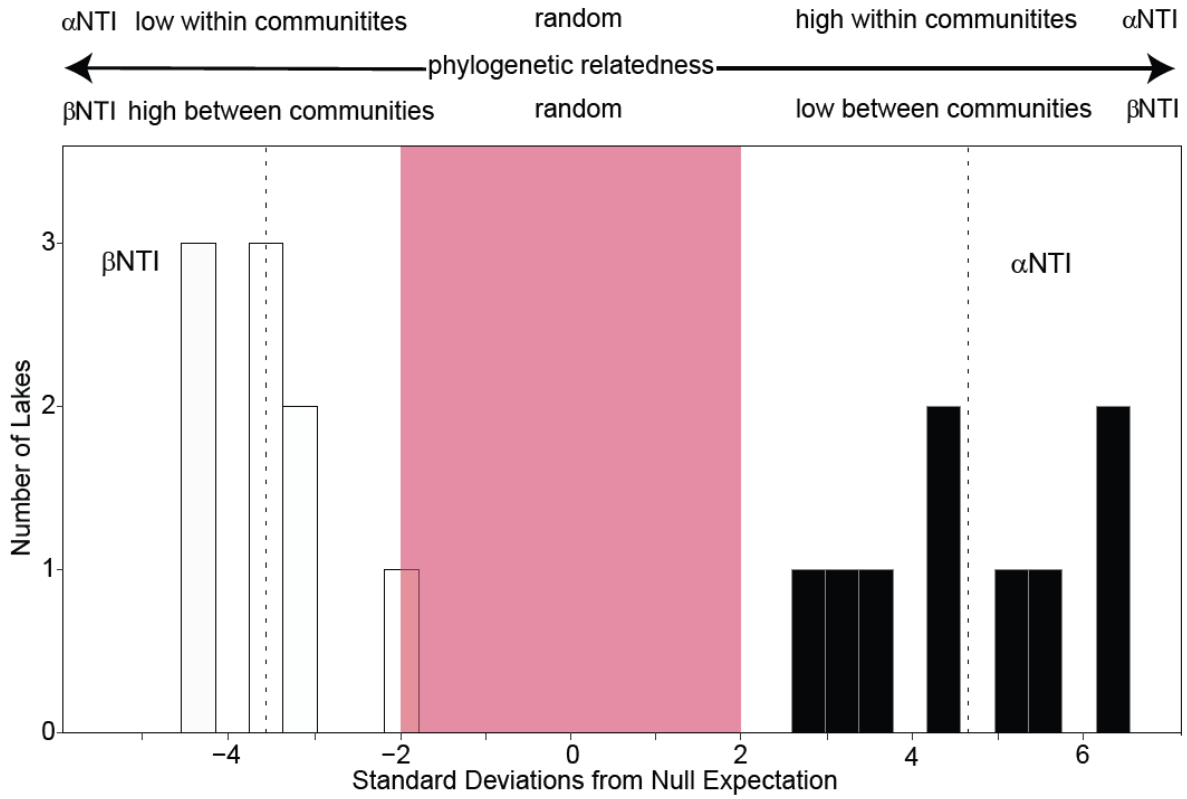


Figure 4. Distributions of *within* community phylogenetic relatedness (αNTI -Nearest Taxon Index), and the phylogenetic relatedness *between* two communities (βNTI) of the nine sampled lakes. Values below -2 or above +2 standard deviations from the null (indicated by the red rectangle) are statistically significantly different from random. Black dashed lines indicate the mean of the observed distributions. The mean of the αNTI distribution is 4.64 and the mean of the βNTI distribution is -3.58. αNTI is a measure of community phylogenetic structure and relatedness, where positive deviations from the null expectation indicate the species in the community are more phylogenetically related (clustered) than expected by chance (as seen here), and negative deviations indicate the species are more phylogenetically distant (over dispersed). The observed αNTI was significantly different from the null ($t = 10.13$, $p < 0.001$). βNTI measures phylogenetic relatedness between two communities with values greater than the null

meaning lower relatedness than expected by chance and values lower than the null meaning higher relatedness than expected by chance (as seen here). Our βNTI is significantly different from random ($t = -12.65$, $p < 0.001$).

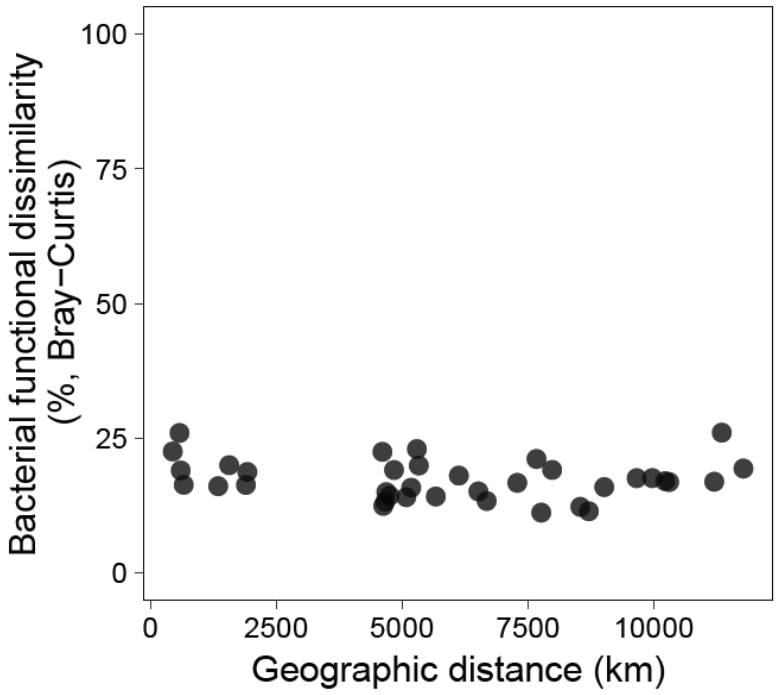


Figure 5. The dissimilarity between the *Microcystis* microbiome community's metagenomic function was not significantly correlated with geographic distance (GLM DE = 1.97%, $p = 0.41$) and was overall low (low values of Bray-Curtis dissimilarity).

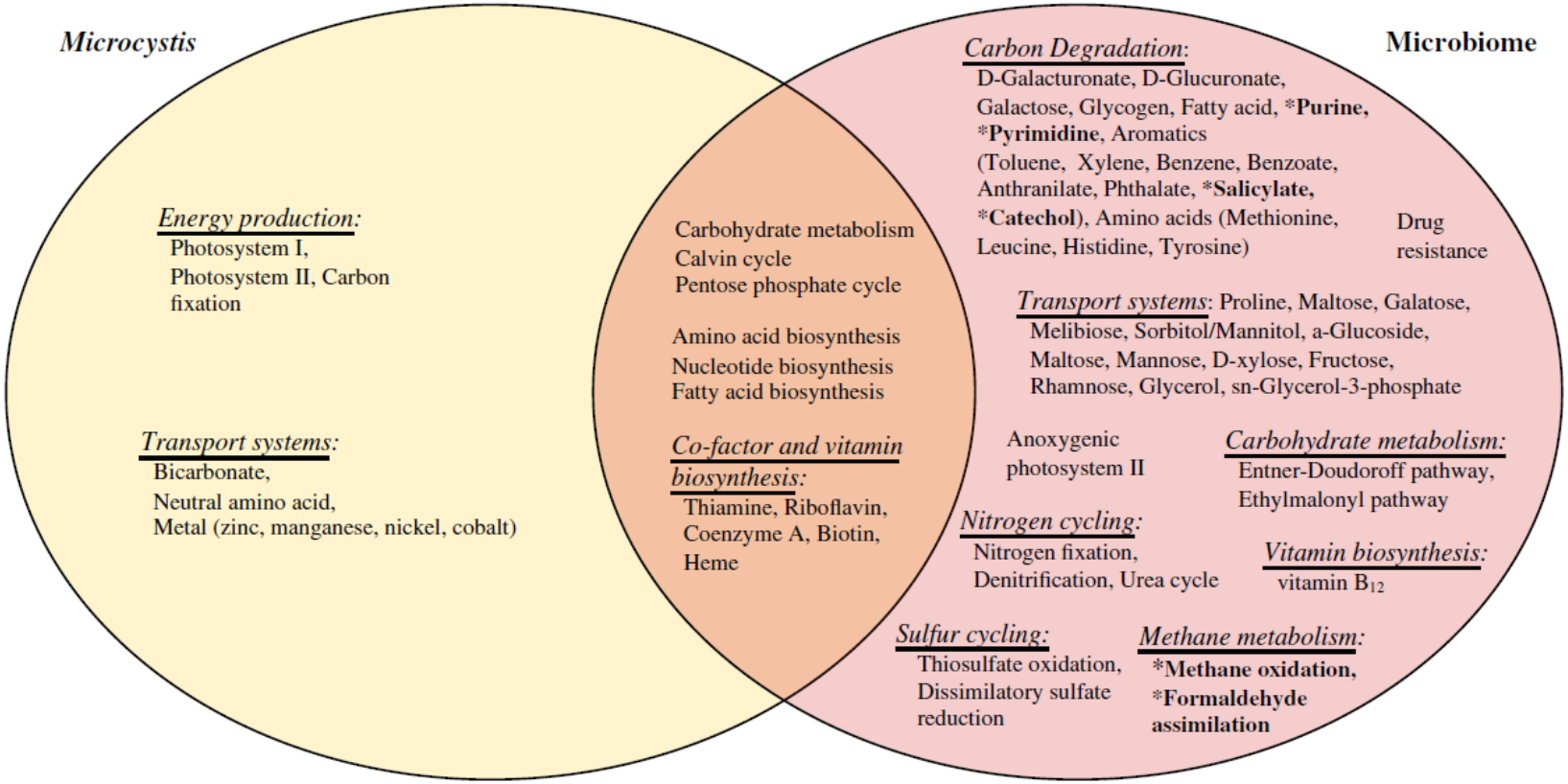


Figure 6. Venn diagram showing the distribution of complete or nearly complete (no more than one gene missing) KEGG modules in Microcystis and the microbiome bacteria. See Supplementary Table 2 for details and indication for involvement in major elemental cycling. KEGG modules in bold print with asterisks were not detected in Microcystis and the microbiome bacterial MAGs.

Chapter 2 - Cyanobacterial bloom succession effects on bacteria community composition over multiple years

Katherine V. Cook,[1,2] J.E. Beyer,[1,2] and K.D. Hambright,[1,2,3]

[1]Plankton Ecology and Limnology Laboratory, Department of Biology, University of Oklahoma, Norman, USA;

[2]Program in Ecology and Evolutionary Biology, Department of Biology, University of Oklahoma, Norman, USA;

[3]Geographical Ecology, Department of Biology, University of Oklahoma, Norman, USA

ABSTRACT

Harmful cyanobacteria blooms are ecosystem disrupting events in freshwater systems that can alter the physico-chemical processes and co-occurring biota, including the associated bacteria community composition. Due to their reduced genomes, cyanobacteria are known to rely on functions provided by bacteria in the surrounding environment. This hypothesized relationship between cyanobacteria and bacteria is called the 'interactome' and we see evidence for it in other cyanoHAB species, such as colonial, aggregate-forming cyanobacteria with mucilaginous sheaths. This morphology allows for close physical proximity with bacteria cells that could facilitate interactions. This concept has not been explored in non-aggregate cyanobacteria, such as *Raphidiopsis*, which are filamentous with no gelatinous sheath and therefore might not be expected to have a microbiome. We hypothesized that if a *Raphidiopsis*-bacteria interactome exists, the bloom would select for a certain bacteria community, resulting in more similar bacterial communities within bloom phases when compared to community composition across bloom phase succession. The hypothesized interactome would cause bacteria communities within the peak of the bloom to be different from non-bloom communities. To test this hypothesis, we sampled a reservoir with seasonally recurrent *Raphidiopsis* blooms for three summers. We collected in-situ environmental data and characterized the microbial community using 16S rRNA sequencing. We found that bacterial communities were more similar between phases year-to-year than between different phases within the same year. The peak bloom bacterial communities were significantly different from pre bloom communities. These results provide initial evidence for a *Raphidiopsis*-bacteria interactome that is conserved over multiple blooms.

INTRODUCTION

Freshwater cyanobacteria blooms can be ecosystem disrupting events, causing changes to biogeochemical cycles, oxygen regimes, and aquatic life (Abdullah Al et al., 2023; Berry et al., 2017; Hans W. Paerl and Otten, 2013; Sunda et al., 2006). Certain cyanobacteria species can also produce toxins and surface scums (Huisman et al., 2018), further impacting water and air quality (Plaas and Paerl, 2021). The primary cause of harmful cyanobacteria blooms (cyanoHABs) is increased eutrophication due to high nutrient loading, specifically nitrogen (N) and phosphorus (P) from anthropogenic sources (Paerl et al., 2016). Most anthropogenic N and P comes from urban and agricultural landscape runoff (Smith and Schindler, 2009), as well as nutrient-rich wastewater discharges (Schindler, 2006). Though recent advancements in nutrient management strategies prioritize the dual reduction of N and P inputs (Paerl et al., 2016), these efforts are not always successful due to internal loading and regeneration of legacy nutrients (Jarvie et al., 2013; Paerl and Barnard, 2020). In addition to persistently high nutrient concentrations, biological interactions can also facilitate cyanoHAB formation.

Among important biological effectors of cyanoHABs are zooplankton and heterotrophic bacteria. Differential grazing by zooplankton has been shown to directly promote cyanobacteria growth in freshwater systems by decreasing competitor biomass (Kibuye et al., 2021). Also, phytoplankton blooms can support substantial heterotrophic bacterial production, mainly as a consequence of increased organic carbon availability (Lignell, 1990). Early research demonstrated not only that phytoplankton are a major DOC source for bacteria, but that associated bacteria can in turn provide key nutrients to phytoplankton via nutrient recycling (Paerl and Pinckney, 1996; Paerl, 1985). In particular, coenobial forms of cyanobacteria like *Microcystis* produce an extracellular polymeric substance (EPS) matrix that provides a unique

habitat for bacteria – the phycosphere – which ensures an intimate cyano-bacterial interaction (Hoke et al., 2021). Indeed, recent research suggests that cyanobacteria-bacteria complementarity may contribute to the success of cyanoHABs (Cai et al., 2014; Cook et al., 2020).

The global success of cyanoHABs is also due to their high growth rate, which is facilitated by their reduced genomes (Cai et al., 2023; Willis et al., 2022). This genome reduction strategy decreases replication costs at the expense of eliminating necessary functions, as described by The Black Queen Hypothesis (Morris et al., 2012). For example, *Raphidiopsis raciborskii* has an incomplete nitrogen-fixation gene cluster (Willis et al., 2022), and *Microcystis* lacks the ability to produce vitamin B_{12} (Li et al., 2018). Because cyanobacteria lack some metabolic pathways necessary for survival, they likely utilize common goods (i.e., metabolites and other necessary resources) produced by bacteria in the surrounding microbiome (Morris et al., 2012). The co-evolution of bacterial interactions could potentially lead to a cyanoHAB 'interactome', a community of synergistic and interacting bacteria species (Garcia et al., 2015). Studies have indeed found that *Microcystis* blooms select for bacteria to account for missing metabolites or functions (Cook et al., 2020; Steffen et al., 2012). While the presence of a *Microcystis* interactome of bacteria attached to the mucilaginous outer layer of its colonies is supported by previous research, non-coenobial cyanobacteria, such as *Raphidiopsis*, which exist as single multicellular trichomes without an EPS matrix (Burford et al., 2016), may not form synergistic relationships with bacteria.

Current knowledge of the bacteria community associated with *Raphidiopsis* blooms and whether they constitute an interactome is limited. Polish lakes with *Raphidiopsis* blooms have significantly different bacterial communities than lakes without (Kokociński et al., 2021). Despite the lack of an EPS matrix, bacteria have been found attached to *Raphidiopsis* cells in

culture and bacterial community composition changed over the course of a bloom in these cultures (Bagatini et al., 2014). Recently, Ou-yang et al. (2022) found that bacteria associated with increased *Raphidiopsis* were different from the bacteria community in other times of the year in a natural system, which shows that *Raphidiopsis* influences bacterial community composition. However, no studies have sought to establish the presence of a Raphidiopsis-bacteria interactome that could be predictably facilitating blooms.

In this study, we sought evidence for the possible existence of a *Raphidiopsis*-bacteria interactome. We hypothesized that if the interactome exists, then *Raphidiopsis* would be a consistently selective force year-to-year, resulting in greater similarity in bacterial microbiome community composition within bloom phases than across bloom phases. Conservation of microbiome composition within bloom phases across multiple years would provide strong evidence for a *Raphidiopsis*-bacteria interactome. In addition, we hypothesized that in the presence of an interactome, the bacterial microbiome would be under the highest selective pressure during the peak of the *Raphidiopsis* bloom, resulting in highly similar peak-bloom communities year-to-year. To further test for the presence of an interactome, we also assessed differences in microbiome composition between different phases of the bloom. Specifically, we predicted that the peak-bloom communities would be significantly different from the pre-bloom and post-bloom communities due to the lack of cyanobacterial influence in the latter two. To test these predictions, we examined bacterial community composition across three years of *Raphidiopsis* blooms in a eutrophic reservoir. Our results indicate that bacteria communities were more similar to other communities in the same phase than communities within the same year and that peak bloom communities are significantly different from pre bloom communities, providing evidence for the existence of a *Raphidiopsis*-bacteria interactome.

METHODS

Study sites and field sampling:

We studied Lake Thunderbird, Oklahoma (OK), a hypereutrophic reservoir with seasonally recurrent blooms of *Raphidiopsis*. These blooms pose challenges to management because the lake services local recreators and provides drinking water to three large, metropolitan areas (Oklahoma Water Resources Board, 2021). We sampled two centrally located near-shore sites including Sailboat (at Little River Marina; sampled from floating doc) and Fisherman's Point (near the municipal water intake; sampled from fixed fishing pier) (Fig. 1). These sites were selected to minimize differences in in depth, limnological parameters, and accessibility. Fisherman's Point (FP) is on the north side of the lake with primary inputs from the Little River, has little cover, and is regularly impacted by strong southerly winds. Sailboat (SB) is also on the north side but faces east, on the Hog Creek arm of the lake, and is in a more protected cove.

We sampled seasonal blooms for three years during the summers of 2016, 2017, and 2018. In 2016, we sampled every day from the 21st of June to the 5th of August. After analyzing data from 2016, we adapted our temporal sampling design to prioritize sampling all bloom phases at the expense of reducing the total number of samples within each phase. Thus, in 2017 we sampled every week from June 15th to July 10th, then sampled every three days after the bloom began until the bloom ended on September 4th. In 2018, we again sampled weekly at the beginning of the season (May 31st – July 3rd) and then approximately every three days once the bloom was detected (July 6th) until the end of the bloom (September 25th).

During each site visit, a Hydrolab DS5X (OTT Hydromet, DE) sonde was used to measure temperature, dissolved oxygen (DO), pH, and phycocyanin. The phycocyanin (PCY) sensor was calibrated using serial dilutions of biologically relevant levels of pure PCY solution (Sigma Aldrich) every year at the beginning of the sampling season. A Kestrel Weather Meter (KestrelMet, USA) was used to measure the max wind speed (m/s), average wind speed (m/s), and air temperature. A LI-COR light meter (LI-COR Biosciences, USA) was used to obtain a light profile to determine the photic zone calculated to the depth where 1.0% of surface incident light reaches. If the light meter was not functioning, we estimated the photic zone as 2.5 times the Secchi depth. This depth was used to calculate photic zone means of environmental variables.

Depth-integrated water samples were collected for DNA extraction, nutrient analysis, and pigment extraction. We used a sampling tube constructed from PVC to ensure each sample pull was collected over the entire photic zone. Multiple pulls were taken to obtain a composite 1-L water sample that was stored in an acid-washed, deionized- and sample-rinsed Nalgene bottle. Samples were immediately placed in a cool, dark container for transport to the laboratory. One 50-mL subsample per analyte was filtered onto separate 25-mm-diameter GF/F filters and stored at –20˚C for subsequent chlorophyll-*a* analyses, and at –80˚C for DNA analysis.

Sample processing:

We determined chlorophyll-*a* (µg/L) using the acetone extraction method following EPA protocol on a Trilogy bench-top fluorometer (see detailed methods: EPA, 1997, section 4.2, Hambright et al. 2014). Nutrient analyses for total nitrogen (TN) and total phosphorus (TP) were performed using EPA QA/QC standards and APHA protocols (APHA 2005) on a Lachat QuikChem FIA (Hach, Loveland, Colorado, USA).

Because we were unable to sequence every sample collected during the study (n=186) to determine bacterial community composition, we selected one sample from each bloom phase every year for DNA extraction and sequencing to determine bacterial community composition. Bloom phases (pre-, log-, peak-, crash-, and post-bloom) were identified by standardizing phycocyanin time-series by year (z-scores) and assigned phases based on the standardized variation of each year's bloom (Fig. S1). This technique allowed us to consistently assign bloom phases across years that differed in bloom magnitude. The last year of the study, 2018, displayed two bloom peaks which resulted in 7 samples per site, while 2016 and 2017 each displayed a single bloom (5 samples per site).

DNA was extracted from the selected filters using the Zymo Quick-DNA Fecal/Soil Microbe Kit (Zymo Research) following the manufacture's recommended protocol. Before placing the GF/F filter in the lysis tube, we cut it up into strips using scissors sterilized between each sample because we found this increased filter degradation during bead lysis. We used the Earth Microbiome Project prokaryote primers 515F (barcoded) and 926R that target the V4-V5 region of the 16S SSU rRNA, which are optimal for profiling both bacterial and archaeal communities by reducing the amplification bias previously seen in some aquatic taxa (Caporaso et al., 2011; Fadeev et al., 2021; Parada et al., 2016; Walters et al., 2015). Each 25-µL PCR reaction mix contained 12.5 µL of 2x DreamTaq Master Mix (ThermoFisher), 1 µL reverse primer (5 µM), 1 µL of the barcoded forward primer (5 µM), 8.5 µL ultrapure water, and 2 µL of DNA. PCR amplification conditions were as follows: Initial denaturation at 98 °C for 3 min, 35 cycles of denaturation at 98 °C for 45 s, annealing at 55 °C for 30 s, extension at 72 °C for 1 min, and final annealing at 72 °C for 10 min. PCR products were visualized on a 1.5% agarose gel to

confirm amplification and amplification quality by the presence of a roughly 480 bp amplicon (amplicon length plus primer construct and Illumina adapter).

PCR products were pooled and then purified using SeraMag magnetic beads (Cytiva). In short, 1.8 × the pooled product volume of the beads was mixed with the pooled PCR product. The solution was then vortexed and incubated at room temperature for 5 min for DNA binding. The 2-mL tube was then placed on a magnetic rack until the solution was clear and beads were visibly clumped by the magnet (~ 5 minutes). Supernatant was removed with a pipette, taking care not to disturb the beads which would result in DNA loss. The beads were then rinsed twice with 80% ethanol and air dried for 5 min. DNA was eluted from beads using 20-uL ultrapure water. Cleaned product was visualized on 1.5% agarose gel to ensure all primers had been removed, and DNA concentration was quantified using a Qubit with the dsDNA High-sensitivity Assay. Once quantification confirmed the product was at least 20-nM, the library was sent for sequencing.

Sequences processing:

Paired-end sequencing for the library was performed on a MiSeq using 2×300 bp v3 chemistry at the Oklahoma Medical Research Foundation (OMRF). Data were demultiplexed and returned via secure ftp for analysis. The 16S raw sequences were processed using Qiime2 (Bolyen et al., 2019). Paired-end reads were not joined, the R2 sequences were discarded due to quality issues, a well-known problem caused by MiSeq 2x300 chemistry. R1 sequences were used for amplicon sequence variant (ASV) identification using the DADA2 ASV algorithm which generates a representative set of high-quality ASV sequences. Taxonomic annotations were assigned to each ASV in Qiime2 using the SILVA SSUv138-99 classifier (Robeson II et al., 2021) using the default confidence threshold of 70% (Bokulich et al., 2018). ASVs that failed

in alignment or were singletons were discarded from the ASV table. To analyze the bacteria community present across the bloom phases we separated the bacteria sequences from cyanobacteria sequences by filtering out all sequences of class *Cyanobacteria*. The combined and bacteria-only ASV abundance tables were used for downstream statistical analyses.

Non-abundant ASVs, those that never exceeded 0.1% in a single sample, and those classified as either eukaryote, archaea, chloroplast, or mitochondria were removed before further analysis. The resulting ASV tables comprised 95% of the sequences. The tables were normalized using Cumulative Sum Scaling (CSS) which corrects for difference in sampling depth across samples (Weiss et al., 2017). After normalizing, there were 430,866 high quality sequences split into 1141 unique ASVs, 1000 of which were bacteria.

Data analyses:

To examine how pigment and nutrient concentrations varied over the course of each bloom, we plotted time-series of each analyte and the nitrogen to phosphorus molar ratios by year. In addition, to display changes in cyanobacterial dominance, we plotted relative abundance of the nine most abundant genera (those displaying relative abundance >55). The samples were aligned on the x-axis based on sampling date to better show how close communities were in time. We show an additional stacked bar plot for only bacterial classes to examine bacterial composition and relative abundances over time. Lastly, to determine how bacteria and cyanobacteria sequences changed over time in relation to each other, we plotted total cyanobacteria sequences and total bacteria sequences on a timeseries.

We examined how bacterial communities associated with *Raphidiopsis* blooms changed during bloom succession by visualizing Bray-Curtis dissimilarity between samples using non-metric multidimensional scaling (NMDS) ordinations of log-transformed ASV sequence counts

(Kruskal, 1964). The final solution was three dimensions, as higher dimensionality only marginally reduced stress. We coded each bacterial community by both year (shape) and bloom phase (color). Because the bacteria are likely influenced by environmental factors distinct from bloom phase, we fit vectors of environmental variables known to influence bacterial communities to the NMDS ordination using the `envfit` function (vegan), including max wind speed (m/s), average wind speed (m/s), air temperature (°C), turbidity (NTU), TN (µg/L), TP (µg/L), TN:TP, water temperature (°C), dissolved oxygen (mg/L), pH, extracted chlorophyll-a (µg/L), and extracted phycocyanin (µg/L). Significant vectors ($p < 0.05$) were displayed on the NMDS ordination, with the length of the environmental vector showing the strength of the correlation.

To determine if similar bacterial communities were present year-to-year during recurrent *Raphidiopsis* blooms, we examined whether bacterial communities of the same bloom phase were more similar to each other (inter-annual similarity between phases) than communities within the same sampling year (intra-annual similarity). We used the distance metric from the NMDS above, but calculated pair-wise comparisons between specific groups to test our hypotheses. To not introduce a spatial effect, sites were only compared with itself, i.e., Fisherman's Point samples compared to other Fisherman's Point samples and Sailboat samples compared to other Sailboat samples. Two groups were created, the first being within phase community comparisons that controlled for year (e.g., if multiple bloom phases occurred in the same year these comparisons were not considered) ($n = 45$) and within year community comparisons for the three years ($n = 82$). We tested for significant differences between these two groups using a Welch's t-test. To further examine how similar the communities where within bloom phase, the Bray-Curtis values from the within phase comparisons ($n = 45$) were grouped

into the individual bloom phases: pre (n = 12), log (n = 6), peak (n = 16), crash (n = 9), and post (n = 2). After removing the post-bloom data due to extremely low sample size, an ANOVA and post-hoc TukeyHSD test were used to determine which bloom phases were significantly more similar to themselves than other bloom phases.

To determine if the peak bloom bacteria community composition was significantly different from the pre and post bloom bacteria communities, we used a permutational multivariate analysis of variance (PERMANOVA) (Anderson, 2001). PERMANOVA tests are sensitive to differences in group dispersion, so first we tested for dispersion differences between the groups using the analysis of multivariate homogeneity (PERMDISP, Anderson, 2006) with the 'betadisper' function. We found a significant dispersion effect ($p = 0.04$) but when we removed the post-bloom samples the dispersion differences were no longer significant. This is likely due to the extremely small sample size we have for the post-bloom sample, which was excluded from the PERMANOVA analysis. However, we were still able to use the 'adonis' function (vegan) to implement a PERMANOVA comparing the pre- and the peak-bloom communities.

To asses potential α-diversity difference caused by the bloom, we used the function 'diversity' from the R package vegan (vegan package v.2.6-4) (Oksanen et al., 2013) to calculate the Shannon Diversity (richness and evenness) and Simpson's D, which takes into account species abundance and ranges from 0 to 1, α-diversity metrics. Evenness, which also ranges from 0 to 1 with higher numbers indicating higher evenness (many species with similar abundances), was calculated as Shannon diversity/log2(number of observed ASVs)). Statistical analyses were completed in the R statistical environment v.4.2.2 (R Development Core Team, 2014).

RESULTS

The chlorophyll-*a* and phycocyanin time series revealed that Lake Thunderbird experienced cyanobacteria blooms of varying magnitudes in all three years of the study (Fig. 2). In 2016, pigment concentrations increased from mid-June to mid-July, with phycocyanin concentrations reaching over 100 µg/L during peak bloom (Fig. 2A & D). Pigment concentrations were still relatively high at the end of sampling, indicating we did not capture the post bloom phase in 2016. We observed the largest bloom of the study the following year in 2017, and measured phycocyanin peaking above 150 µg/L in August (Fig. 2E). In 2018, there was a smaller, more gradual bloom that increased to a first peak in mid-July before crashing at the beginning of August. There was a second peak in September that was higher than the first, with phycocyanin reaching 88 µg/L (Fig. 2F). We began sampling earlier in 2018, and with the prolonged bloom, we were able to capture two pre-bloom samples and two peak samples (Fig. S1).

Nutrient concentrations ranged between 18.9 – 71.0 µg/L of TP and 172 – 1430 µg/L of TN over the course of the study (Table 1). TP and TN generally increased with the rise of the bloom (Fig. S2). The highest concentrations of TN occurred in 2017 during the crash phase of the bloom, whereas maximum TP occurred during the post-bloom phase in 2018 (Fig. S2B & F). N to P ratios were relatively low throughout the study (Fig. S5). We saw the lowest ratios early in the season, in the pre bloom phases, and in crash and post phases of 2018. Based on Dzialowski et al.'s (2005) classification of nutrient limitation in other Great Plains reservoirs, the low ratio indicated the system might be experiencing nitrogen limitation during these times. The majority of the time, our samples fell in the nitrogen and phosphorus co-limiting range (between 18-45), but in practice, limitation is unlikely due to the high concentrations of nutrients in the

water (Fig. S2). Water temperature ranged from 24 – 31 °C and was driven by a combination of warm air temperatures (mean 30 °C) and shallow, well-mixed sites. We saw very little hypoxia and in general the DO was high (mean 7.3 mg/L) and evenly distributed throughout the water column. Even with high winds, our sites had relatively low turbidity (mean 8.4 NTU) compared to other places on Lake Thunderbird that regularly reach over 50 NTU (pers. observation).

Based on cursory light microscopy at the time of sampling, we know the main bloom forming cyanobacteria in Lake Thunderbird is *Raphidiopsis* sp. (previously *Cylindrospermopsis*), and sequencing results corroborated this finding (Fig. 3). In most cases, on the dates corresponding with high pigment values, *Raphidiopsis* was the dominate cyanobacteria (Fig. 3, 2016 and 2017). In one instance, during the second peak phase sample in 2016 at Fisherman's Point (see Fig. S1 for phase designations), the low cyanobacteria sequence count did not accurately reflect the very high phycocyanin and chlorophyll-*a* concentrations, which indicate the bloom was still abundant (Fig. 3) and could indicate differential amplification or low sequencing depth in this sample. There was some seasonal variation in cyanobacteria dominance, with *Cyanobium*, a picocyanobacteria that typically occurs as solidary cells (Jasser and Callieri, 2016), being consistently more abundant in the pre-bloom phase across all three years. In 2018, the bloom was more of a mixed-species bloom than previous years, with *Cyanobium*, *Prochlorothrix,* a small, filamentous cyanobacteria, and *Leptolyngbya* (*Nodosilinea*), another filamentous cyanobacterium.

Bacteria sequence abundance varied annually and roughly correlated with bloom succession (Fig. 4, Fig. S4). The highest total abundance occurred in 2016 at Sailboat with around 20,000 bacteria sequences. *Verrucomicrobiae, Bacteroidia, Gammaproteobacteria,* and *Actinobacteria* where the dominant bacteria classes throughout each year (Fig. 4). Other classes,

including *Alphaproteobacteria*, *Planctomycetes*, *Phycisphaerae*, and *Anaerolineae*, were also present at lower abundances (Fig. 4). Bacterial sequence abundance generally tracked cyanobacterial sequence abundance (Fig. S4), with the exception of 2016 where in the crash phases the bacteria sequences were substantially higher than the cyanobacteria.

Overall, alpha diversity of the bacteria community did not vary over the course of the bloom or between years (Fig. S3). Shannon Diversity was high with no consistent trend over time (Fig. S3A & B). Simpson's D was very high over the whole study (Fig. S3E & F), indicating high diversity. Evenness was also high throughout the study, indicating more evenly distributed species and abundances (Fig. S3C & D). Lastly, the observed number of ASVs did vary over time, again with no discernible trend with bloom phase. For example, in 2018 the highest number of bacteria species occurred in the pre-bloom phases, while in 2016 the highest number of bacteria species occurred in the peak-bloom phase (Fig. S3G & H).

Non-metric multidimensional scaling (NMDS) based on dissimilarities in 16S bacterial community composition revealed the greatest separation between the pre-bloom and other bloom phases in ordination space (Fig. 5A & B), which shows pre-bloom communities were similar to each other and distinct from communities in other bloom phases. The phases within the bloom (log, peak, and crash) displayed considerable overlap in ordination space, indicating increased community similarity in the presence of the *Raphidiopsis* bloom. Environmental variables where fit as vectors onto the NMDS to explore the relationship between bacteria community composition and the local environment. We found chlorophyll-*a*, phycocyanin, TP, TN, and N:P were significant environmental correlates to differences in bacterial community composition (Fig. 5A, Table 2). Specifically, the separation of bloom phases, or at least the pre-bloom phase from all other phases, aligned with the chlorophyll-*a* and phycocyanin vectors on NMDS1. This

indicates the *Raphidiopsis* bloom had a hand in structuring the bacteria composition of these communities. Nutrient concentrations were also correlated with community separation along NMDS2, which shows some separation by year and phase, specifically for 2018 (triangles) which had a large TP gradient. Variation in bacterial community composition also correlated with water temperature, air temperature, DO, pH, turbidity, max wind speed, and average wind speed when fit against NMDS1 and NMDS3 (Fig 5B, Table 2).

To test our main hypothesis, that there is greater inter-annual similarity among bloom phase than intra-annual similarity within years, we compared the community dissimilarity among bloom phases across years (pre, log, peak, crash, post) and within years (2016, 2017, 2018). We found that bacterial communities associated with *Raphidiopsis* blooms were more similar between phases year-to-year than between different phases across the same year (Fig 6A; Welch's t-test: t = -2.3157, df = 101.4, p-value = 0.02), which indicates that bloom succession was influencing bacterial community composition in similar ways across years. To further explore the similarities within phases, we examined community dissimilarity values within each phase (Fig. 6B). We found there was a significant difference between similarity within bloom phases (ANOVA, df =3, SumSq = 0.06, MeanSq = 0.02, F = 4.2, p = 0.01) and upon post-hoc testing determined this difference was between the peak and crash communities (TukeyHSD p = 0.007). This indicates that bacteria community composition at the peak of the bloom were quite similar year-to-year, while crash community composition was much more stochastic. Based on this analysis (Fig. 6B), the pre, log, and peak communities all had indistinguishable within-phase community similarity, but visually the peak bloom phase had the highest community similarity. Post-bloom samples were not included in this analysis due to extreme sample size differences, but based on visual inspection, the post bloom communities were very different from each other

(high Bray-Curtis values) and more dissimilar than the other communities (Fig. 6B). Next we tested if the peak phase bacteria communities composition were significantly different from the pre phase bacteria communities composition, and we found that the bacteria community compositions were significantly different from each other (PERMANOVA: df = 1, SumOfSqs = 0.78, R2 = 0.3, F = 7.5, p = 0.001), confirming our hypothesis that the cyanobacteria bloom's selection of bacterial community composition is conserved over multiple years and corroborating the existence of a *Raphidiopsis* interactome.

DISCUSSION

Cyanobacteria blooms are increasing globally due to climate change and pervasive nutrient pollution, and can disrupt ecosystems by altering environmental conditions and microbial community composition (Bao et al., 2023; Berry et al., 2017; Huisman et al., 2018). Cyanobacteria accrue a fitness advantage by employing a genome reduction strategy commonly seen in bacteria, but at the expense of losing necessary metabolic functions (Morris et al., 2011; Willis et al., 2022, 2018). They likely rely on the surrounding community, mostly bacteria, to fulfill these metabolic needs, leading to interactions with select bacteria, i.e., an 'interactome' (Li et al., 2018; Xie et al., 2016). We hypothesized that *Raphidiopsis* blooms select for certain bacteria and this bacterial microbiome is conserved across multiple years.

Genome sequencing confirmed that *Raphidiopsis* was the most abundant cyanobacteria in Lake Thunderbird, but that bloom peaks varied in timing and magnitude each year. Other cyanobacteria, such as *Cyanobium*, *Leptolyngbya* (*Nodosilinea*), *Prochlorothrix,* and *Cuspidothrix,* occurred at low levels during the bloom. *Cyanobium* was seen at higher amounts during the low *Raphidiopsis* periods early in the season and made up a large proportion of the

pre-bloom cyanobacteria community, even though the pigments in the pre-bloom phases were low. This is likely due to the small size of *Cyanobium* cells and therefore low overall biomass relative to genomic material. This transition in cyanobacterial dominance from early season *Cyanobium* to a summer *Raphidiopsis* bloom is typical for Lake Thunderbird (per. communication with Haiyuan Cai) and has been seen in other systems. A study on Utah Lake found high *Cyanobium* abundance in the early summer from May to July before transitioning into a different filamentous cyanobacterial (*Aphanizomenon* and *Dolichospermum)*, bloom (Li et al., 2020). *Cyanobium* is in the order Synechocaccales, a part of the picocyanobacterial clade, and is likely a dominant competitor due to fast nutrient uptake and growth rates made possible by its small cellular volume and larger surface-to-volume ratio (Suttle and Harrison, 1986). It is not, however, capable of fixing nitrogen, and therefore could be outcompeted by the nitrogen-fixer *Raphidiopsis*. We saw low N:P ratios in the pre-bloom sample (< 18 N:P), which might indicate that lower nitrogen concentrations relative to phosphorus facilitate bloom formation in Lake Thunderbird.

Bacteria previously associated with a *Raphidiopsis* bloom (Ou-yang et al., 2022), were present throughout our study, with *Verrucomicrobiae, Bacteroidia, Gammaproteobacteria,* and *Actinobacteria* representing the majority of the bacteria sequences (Eiler and Bertilsson, 2004; Woodhouse et al., 2016). Generally, bacterial abundances increased as cyanobacteria increased, with the exception of 2016, where bacterial abundances were highest during the crash phases of the bloom at Sailboat. Bacteria abundance changes mirroring those in the cyanobacteria community has been seen in other systems (Ou-yang et al., 2022), indicating an overall increase in lake productivity at the height of the bloom. Higher bacteria abundance has also been recorded at the death of blooms (Bouvy et al., 2001). This aligns with observations that large amounts of

organic matter are present during blooms (Bai et al., 2017; Bouvy et al., 2001) and released in large quantities upon cyanobacteria cell death (Leloup et al., 2013; H. W. Paerl and Otten, 2013; Ye et al., 2012).

We found evidence for the existence of a *Raphidiopsis*-bacteria interactome. Specifically, we tested the hypothesis that *Raphidiopsis* blooms would be a structuring force for the bacterial microbiome, resulting in more similar bacterial communities within bloom phases then within bloom phases in a given year. We found that bacterial communities from distinct bloom phases (pre, log, peak, crash, and post) were more similar to communities within those phases across multiple years than they were to the other communities within their same year. As we see taxonomically conserved bacterial microbiome succession across multiple years, this indicates that the *Raphidiopsis* interactome could be strongly shaping bacterial community composition. We also found significant differences in bacterial community composition between the pre bloom and peak bloom phases. This supports our hypothesis that in the presence of strong selection by *Raphidiopsis*, the peak communities would be distinctly different from communities lacking *Raphidiopsis*. It is, however, interesting that the pre bloom communities are very similar to each other. There could be a different selection force, such as spring zooplankton community or cooler temperatures, that are causing those communities to converge on a similar bacteria community every year.

In addition, our study showed that bacteria communities are most similar year-to-year during the peak phase and that the following crash introduces significant differences in community composition when compared to the peak phase. This indicates *Raphidiopsis* selective pressure is strongest at highest abundances (i.e., the peak of the bloom) and adds to the body of evidence for a *Raphidiopsis*-bacteria interactome. The lessening of selective pressure of

Raphidiopsis as abundances decrease likely contributes to the crash communities' increased dissimilarity compared to the peak phase. Other studies have shown a dramatic increase in predatory bacteria at the decline of a cyanobacteria bloom (Rashidan and Bird, 2001), and we speculate crash and post-bloom dissimilarity is caused by a shift in community structuring processes, potentially from mutualistic *Raphidiopsis* selection to competition or predation in the bacteria community.

This pattern of conserved changes in bacterial community composition year to year could alternatively be explained by changes in the local environment. Bacteria community assembly and composition are affected by abiotic factors, for example studies have shown water temperature (Kan et al., 2006; Kritzberg et al., 2006), dissolved oxygen, nitrate nitrogen (Su et al., 2017), and dissolved organic carbon source (Kritzberg et al., 2006) all impact bacteria community composition. In our study, we too found that water temperature, nitrogen and phosphorus concentrations, and dissolved oxygen impacted bacteria community composition, in addition to other environmental factors. Cyanobacteria can alter the surrounding local environment over the course of the bloom by generating abundant dissolved organic carbon, as previously mentioned, changing dissolved oxygen regimes (Paerl, 2022; Wang and Zhang, 2020), and altering nutrient cycles (Burford et al. 2016). If the *Raphidiopsis* bloom predictably changes organic matter and nutrient cycles every bloom, this could lead to similar bacterial communities year to year. Additional data is needed to elucidate the mechanism conserving bacterial community composition, including dissolved organic matter concentration and types, as well as the distribution of microbial pathways and functions in the interactome community.

Although cyanobacteria-bacteria interactions are probable due to *Raphidiopsis's* reduced genome, functional analyses of the community are necessary to determine this. Studies have

shown that cyanobacteria and their bacteria microbiome contribute unique pathways to the total metabolic function in the community (Cook et al., 2020; Shi et al., 2022). Specifically, Shi (2022) found *Microcystis* did not have all necessary genes for the nitrogen or sulfur metabolic pathways and those needed genes where present in a number of abundant bacteria in its microbiome. Considering some *Raphidiopsis* species have been found with an incomplete nitrogen-fixation gene cluster (Willis et al., 2022), this same pattern could be present in a *Raphidiopsis*-bacteria interactome. Including functional analyses through metagenomic sequencing in future studies would help determine if distinct interactions are occurring between the cyanobacterium and associated bacteria.

The observed bacteria community changes in response to *Raphidiopsis* are consistent with prior studies (Kokociński et al., 2021; Ou-yang et al., 2022), but to our knowledge no other study has examined temporally reoccurring *Raphidiopsis* blooms. Indeed, in cyanobacteria-bacteria research, temporal studies lasting longer than a year are rare. Most studies investigating the bacteria community associated with other species of phytoplankton blooms do not explore recurring trends in bacterial community composition over multiple years (Bagatini et al., 2014; Berry et al., 2017; Cai et al., 2014; Lezcano et al., 2017; Li et al., 2020; Niu et al., 2011; Ou-yang et al., 2022; Su et al., 2017; Teeling et al., 2012; Woodhouse et al., 2016). Shi et al. (2022) did study the bacteria associated with *Microcystis* in Taihu over the course of 16 months, but only captured a single bloom period. Tromas et al. (2017) combined an impressive data set with eight years of microbial community data from blooms on Lake Champlain to study bacterial community succession from non-bloom to bloom (*Microcystis* and *Dolichospermum*) time periods and found a recurring and significant difference between non-bloom and bloom bacteria communities. Multi-annual studies examining only bacterioplankton community composition are

also limited in number (but see Kan et al., 2006; Kent et al., 2004). Future temporal studies of cyanobacteria-bacteria interactomes should attempt to capture more than one bloom period to help determine whether selection of bacteria is conserved over multiple blooms and multiple years. Long-term temporal studies would help elucidate if the interactome are in fact co-evolved and conserved through time.

In this study, we used 16S taxonomy of the microbial community from cyanobacteria blooms in three years to seek evidence for the existence of a *Raphidiopsis*-bacteria interactome. We tested the hypothesis that if *Raphidiopsis* blooms were a structuring force for the bacterial microbiome, resulting bacterial communities would be more similar within bloom phases when compared to community composition across bloom phase succession. We found that bacterial communities associated with *Raphidiopsis* blooms were more similar within phase year-to-year than between different phases across the same year. These results demonstrate that *Raphidiopsis* blooms are a major selection factor for the bacteria community within lake ecosystems. This is a first step in providing evidence for the existence of a co-evolved *Raphidiopsis*-bacteria interactome and these results highlight the need to further investigation into the temporal dynamics of *Raphidiopsis*-bacteria interactions and how community function may play a role in bacterial community composition.

ACKNOWLEDGMENTS

We thank the members of the University of Oklahoma Plankton Ecology and Limnology Laboratory for helpful discussions, field sampling assistance, and lab assistance, particularly Thayer Hallidayschult, Haiyuan Cai, and Chris McLimans. We thank Stephen Cook for his invaluable statistical advice and critical editorial feedback. We appreciate Jeff Back, CRASR

Baylor University, for processing and discussion about our nutrient samples. Funding was provided by the University of Oklahoma Office of the Vice President for Research to L.R.K. and K.D.H., by the National Science Foundation (DEB 1831061) to K.D.H., L.R.K., H.W.P., M.M.S., and A.E.W., NSF OCE 1840715 and NIH-1P01ES028939-01 to H.W.P, NSF DEB 1831106 to M.M.S., and through an NSF Graduate Research Fellowship and University of Oklahoma Nancy L. Mergler Dissertation Completion Fellowship to K.V.C.

WORKS CITED

Abdullah Al, M., Wang, W., Jin, L., Chen, H., Xue, Y., Jeppesen, E., Majaneva, M., Xu, H., Yang, J., 2023. Planktonic ciliate community driven by environmental variables and cyanobacterial blooms: A 9-year study in two subtropical reservoirs. Science of The Total Environment 858, 159866. https://doi.org/10.1016/j.scitotenv.2022.159866

Anderson, M.J., 2006. Distance-Based Tests for Homogeneity of Multivariate Dispersions. Biometrics 62, 245–253. https://doi.org/10.1111/j.1541-0420.2005.00440.x

Anderson, M.J., 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology 26, 32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

Antunes, J.T., Leão, P.N., Vasconcelos, V.M., 2015. <i>Cylindrospermopsis raciborskii<i>: review of the distribution, phylogeography, and ecophysiology of a global invasive species. Front. Microbiol. 0. https://doi.org/10.3389/fmicb.2015.00473

Bagatini, I.L., Eiler, A., Bertilsson, S., Klaveness, D., Tessarolli, L.P., Vieira, A.A.H., 2014. Host-Specificity and Dynamics in Bacterial Communities Associated with Bloom-Forming Freshwater Phytoplankton. PLoS One 9, e85950. https://doi.org/10.1371/journal.pone.0085950

Bai, L., Cao, C., Wang, C., Xu, H., Zhang, H., Slaveykova, V.I., Jiang, H., 2017. Toward Quantitative Understanding of the Bioavailability of Dissolved Organic Matter in Freshwater Lake during Cyanobacteria Blooming. Environ. Sci. Technol. 51, 6018–6026. https://doi.org/10.1021/acs.est.7b00826

Bao, Y., Huang, T., Ning, C., Sun, T., Tao, P., Wang, J., Sun, Q., 2023. Changes of DOM and its correlation with internal nutrient release during cyanobacterial growth and decline in Lake Chaohu, China. Journal of Environmental Sciences 124, 769–781. https://doi.org/10.1016/j.jes.2022.02.019

Berry, M.A., Davis, T.W., Cory, R.M., Duhaime, M.B., Johengen, T.H., Kling, G.W., Marino, J.A., Den Uyl, P.A., Gossiaux, D., Dick, G.J., Denef, V.J., 2017. Cyanobacterial harmful algal blooms are a biological disturbance to Western Lake Erie bacterial communities. Environmental Microbiology 19, 1149–1162. https://doi.org/10.1111/1462-2920.13640

Bokulich, N.A., Kaehler, B.D., Rideout, J.R., Dillon, M., Bolyen, E., Knight, R., Huttley, G.A., Gregory Caporaso, J., 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6, 90. https://doi.org/10.1186/s40168-018-0470-z

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L.,

Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37, 852–857. https://doi.org/10.1038/s41587-019-0209-9

Bouvy, M., Pagano, M., Troussellier, M., 2001. Effects of a cyanobacterial bloom (Cylindrospermopsis raciborskii) on bacteria and zooplankton communities in Ingazeira reservoir (northeast Brazil). Aquat. Microb. Ecol. 25, 215–227. https://doi.org/10.3354/ame025215

Burford, M.A., Beardall, J., Willis, A., Orr, P.T., Magalhaes, V.F., Rangel, L.M., Azevedo, S.M.F.O.E., Neilan, B.A., 2016. Understanding the winning strategies used by the bloom-forming cyanobacterium Cylindrospermopsis raciborskii. Harmful Algae, Global Expansion of Harmful Cyanobacterial Blooms: Diversity, ecology, causes, and controls 54, 44–53. https://doi.org/10.1016/j.hal.2015.10.012

Cai, H., Jiang, H., Krumholz, L.R., Yang, Z., 2014. Bacterial community composition of size-fractioned aggregates within the phycosphere of cyanobacterial blooms in a eutrophic freshwater lake. PLoS One 9, e102879. https://doi.org/10.1371/journal.pone.0102879

Cai, H., McLimans, C.J., Beyer, J.E., Krumholz, L.R., Hambright, K.D., 2023. *Microcystis* pangenome reveals cryptic diversity within and across morphospecies. Sci. Adv. 9, eadd3783. https://doi.org/10.1126/sciadv.add3783

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N., Knight, R., 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. PNAS 108, 4516–4522. https://doi.org/10.1073/pnas.1000080107

Cook, K.V., Li, C., Cai, H., Krumholz, L.R., Hambright, K.D., Paerl, H.W., Steffen, M.M., Wilson, A.E., Burford, M.A., Grossart, H.-P., Hamilton, D.P., Jiang, H., Sukenik, A., Latour, D., Meyer, E.I., Padisák, J., Qin, B., Zamor, R.M., Zhu, G., 2020. The global Microcystis interactome. Limnology and Oceanography 65, S194–S207. https://doi.org/10.1002/lno.11361

Dzialowski, A.R., Wang, S.-H., Lim, N.-C., Spotts, W.W., Huggins, D.G., 2005. Nutrient limitation of phytoplankton growth in central plains reservoirs, USA. Journal of Plankton Research 27, 587–595. https://doi.org/10.1093/plankt/fbi034

Eiler, A., Bertilsson, S., 2004. Composition of freshwater bacterial communities associated with cyanobacterial blooms in four Swedish lakes. Environ. Mircobiol. 6, 1228–1243. https://doi.org/10.1111/j.1462-2920.2004.00657.x

Fadeev, E., Cardozo-Mino, M.G., Rapp, J.Z., Bienhold, C., Salter, I., Salman-Carvalho, V., Molari, M., Tegetmeyer, H.E., Buttigieg, P.L., Boetius, A., 2021. Comparison of Two

16S rRNA Primers (V3–V4 and V4–V5) for Studies of Arctic Microbial Communities. Frontiers in Microbiology 12.

Garcia, S.L., Buck, M., McMahon, K.D., Grossart, H.-P., Eiler, A., Warnecke, F., 2015. Auxotrophy and intrapopulation complementary in the 'interactome' of a cultivated freshwater model community. Mol. Ecol. 24, 4449–4459. https://doi.org/10.1111/mec.13319

Hoke, A.K., Reynoso, G., Smith, M.R., Gardner, M.I., Lockwood, D.J., Gilbert, N.E., Wilhelm, S.W., Becker, I.R., Brennan, G.J., Crider, K.E., Farnan, S.R., Mendoza, V., Poole, A.C., Zimmerman, Z.P., Utz, L.K., Wurch, L.L., Steffen, M.M., 2021. Genomic signatures of Lake Erie bacteria suggest interaction in the Microcystis phycosphere. PLOS ONE 16, e0257017. https://doi.org/10.1371/journal.pone.0257017

Huisman, J., Codd, G.A., Paerl, H.W., Ibelings, B.W., Verspagen, J.M.H., Visser, P.M., 2018. Cyanobacterial blooms. Nat Rev Microbiol 16, 471–483. https://doi.org/10.1038/s41579-018-0040-1

Jarvie, H.P., Sharpley, A.N., Withers, P.J.A., Scott, J.T., Haggard, B.E., Neal, C., 2013. Phosphorus Mitigation to Control River Eutrophication: Murky Waters, Inconvenient Truths, and "Postnormal" Science. Journal of Environmental Quality 42, 295–304. https://doi.org/10.2134/jeq2012.0085

Jasser, I., Callieri, C., 2016. Picocyanobacteria, in: Handbook of Cyanobacterial Monitoring and Cyanotoxin Analysis. John Wiley & Sons, Ltd, pp. 19–27. https://doi.org/10.1002/9781119068761.ch3

Kan, J., Crump, B.C., Wang, K., Chen, F., 2006. Bacterioplankton community in Chesapeake Bay: Predictable or random assemblages. Limnology and Oceanography 51, 2157–2169. https://doi.org/10.4319/lo.2006.51.5.2157

Kent, A.D., Jones, S.E., Yannarell, A.C., Graham, J.M., Lauster, G.H., Kratz, T.K., Triplett, E.W., 2004. Annual Patterns in Bacterioplankton Community Variability in a Humic Lake. Microb Ecol 48, 550–560. https://doi.org/10.1007/s00248-004-0244-y

Kibuye, F.A., Zamyadi, A., Wert, E.C., 2021. A critical review on operation and performance of source water control strategies for cyanobacterial blooms: Part II-mechanical and biological control methods. Harmful Algae 109, 102119. https://doi.org/10.1016/j.hal.2021.102119

Kokociński, M., Dziga, D., Antosiak, A., Soininen, J., 2021. Are Bacterio- and Phytoplankton Community Compositions Related in Lakes Differing in Their Cyanobacteria Contribution and Physico-Chemical Properties? Genes 12, 855. https://doi.org/10.3390/genes12060855

Kritzberg, E.S., Langenheder, S., Lindström, E.S., 2006. Influence of dissolved organic matter source on lake bacterioplankton structure and function – implications for seasonal dynamics of community composition. FEMS Microbiol Ecol 56, 406–417. https://doi.org/10.1111/j.1574-6941.2006.00084.x

Kruskal, J.B., 1964. Nonmetric multidimensional scaling: A numerical method. Psychometrika 29, 115–129. https://doi.org/10.1007/BF02289694

Leloup, M., Nicolau, R., Pallier, V., Yéprémian, C., Feuillade-Cathalifaud, G., 2013. Organic matter produced by algae and cyanobacteria: Quantitative and qualitative

characterization. Journal of Environmental Sciences 25, 1089–1097. https://doi.org/10.1016/S1001-0742(12)60208-3

Lezcano, M.Á., Velázquez, D., Quesada, A., El-Shehawy, R., 2017. Diversity and temporal shifts of the bacterial community associated with a toxic cyanobacterial bloom: An interplay between microcystin producers and degraders. Water Research 125, 52–61. https://doi.org/10.1016/j.watres.2017.08.025

Li, H., Barber, M., Lu, J., Goel, R., 2020. Microbial community successions and their dynamic functions during harmful cyanobacterial blooms in a freshwater lake. Water Research 185, 116292. https://doi.org/10.1016/j.watres.2020.116292

Li, Q., Lin, F., Yang, C., Wang, J., Lin, Y., Shen, M., Park, M.S., Li, T., Zhao, J., 2018. A large-scale comparative metagenomic study reveals the functional interactions in six bloom-forming microcystis-epibiont communities. Front. Microbiol. 9. https://doi.org/10.3389/fmicb.2018.00746

Lignell, R., 1990. Excretion of organic carbon by phytoplankton: its relation to algal biomass, primary productivity and bacterial secondary productivity in the Baltic Sea. Marine Ecology Progress Series 68, 85–99.

Morris, J.J., Johnson, Z.I., Szul, M.J., Keller, M., Zinser, E.R., 2011. Dependence of the cyanobacterium *Prochlorococcus* on hydrogen peroxide scavenging microbes for growth at the ocean's surface. PLoS One 6, e16805. https://doi.org/10.1371/journal.pone.0016805

Morris, J.J., Lenski, R.E., Zinser, E.R., 2012. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. MBio 3, e00036-12.

Niu, Y., Shen, H., Chen, J., Xie, P., Yang, X., Tao, M., Ma, Z., Qi, M., 2011. Phytoplankton community succession shaping bacterioplankton community composition in Lake Taihu, China. Water Research 45, 4169–4182. https://doi.org/10.1016/j.watres.2011.05.022

Oklahoma Water Resources Board, 2021. Lake Thunderbird Water Quality 2021 Final Report (Lake Report). Oklahoma Water Resources Board, State of Oklahoma, Oklahoma City, OK.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'hara, R., Simpson, G.L., Solymos, P., Stevens, M.H.H., Wagner, H., 2013. Package 'vegan.' Community ecology pakage 2, 631–637.

Ou-yang, T., Yang, S., Zhao, L., Ji, L., Shi, J., Wu, Z., 2022. Temporal heterogeneity of bacterial communities and their responses to Raphidiopsis raciborskii blooms. Microbiological Research 262, 127098. https://doi.org/10.1016/j.micres.2022.127098

Paerl, H., Pinckney, J., 1996. A Mini-review of Microbial Consortia: Their Roles in Aquatic Production and Biogeochemical Cycling. https://doi.org/10.1007/BF00171569

Paerl, H.W., 2022. Human and Climatic Drivers of Harmful Cyanobacterial Blooms (CyanoHABs), in: Assessing the Microbiological Health of Ecosystems. John Wiley & Sons, Ltd, pp. 31–48. https://doi.org/10.1002/9781119678304.ch2

Paerl, H.W., 1985. Microzone formation: Its role in the enhancement of aquatic N2 fixation1. Limnology and Oceanography 30, 1246–1252. https://doi.org/10.4319/lo.1985.30.6.1246

Paerl, H.W., Barnard, M.A., 2020. Mitigating the global expansion of harmful cyanobacterial blooms: Moving targets in a human- and climatically-altered world. Harmful Algae 96, 101845. https://doi.org/10.1016/j.hal.2020.101845

Paerl, Hans W., Otten, T.G., 2013. Harmful cyanobacterial blooms: Causes, consequences, and controls. Microb. Ecol. 65, 995–1010. https://doi.org/10.1007/s00248-012-0159-y

Paerl, H. W., Otten, T.G., 2013. Blooms bite the hand that feeds them. Science 342, 432–433. https://doi.org/10.1126/science.1246252

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., 2016. It takes two to tango: When and where dual nutrient (N & P) reductions are needed to protect lakes and downstream ecosystems. Environmental science & technology 50, 10805–10813.

Parada, A.E., Needham, D.M., Fuhrman, J.A., 2016. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environmental Microbiology 18, 1403–1414. https://doi.org/10.1111/1462-2920.13023

Plaas, H.E., Paerl, H.W., 2021. Toxic Cyanobacteria: A Growing Threat to Water and Air Quality. Environ. Sci. Technol. 55, 44–64. https://doi.org/10.1021/acs.est.0c06653

R Development Core Team, 2014. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rashidan, K.K., Bird, D.F., 2001. Role of predatory bacteria in the termination of a cyanobacterial bloom. Microb Ecol 41, 97–105. https://doi.org/10.1007/s002480000074

Robeson II, M.S., O'Rourke, D.R., Kaehler, B.D., Ziemski, M., Dillon, M.R., Foster, J.T., Bokulich, N.A., 2021. RESCRIPt: Reproducible sequence taxonomy reference database management. PLOS Computational Biology 17, e1009581. https://doi.org/10.1371/journal.pcbi.1009581

Sánchez-Baracaldo, P., Hayes, P.K., Blank, C.E., 2005. Morphological and habitat evolution in the Cyanobacteria using a compartmentalization approach. Geobiology 3, 145–165. https://doi.org/10.1111/j.1472-4669.2005.00050.x

Schindler, D.W., 2006. Recent advances in the understanding and management of eutrophication. Limnol. Oceanogr. 51, 356–363. https://doi.org/10.4319/lo.2006.51.1_part_2.0356

Shi, L., Cai, Y., Gao, S., Zhang, M., Chen, F., Shi, X., Yu, Y., Lu, Y., Wu, Q.L., 2022. Gene expression pattern of microbes associated with large cyanobacterial colonies for a whole year in Lake Taihu. Water Research 223, 118958. https://doi.org/10.1016/j.watres.2022.118958

Smith, V.H., Schindler, D.W., 2009. Eutrophication science: where do we go from here? Trends in Ecology & Evolution 24, 201–207. https://doi.org/10.1016/j.tree.2008.11.009

Steffen, M.M., Li, Z., Effler, T.C., Hauser, L.J., Boyer, G.L., Wilhelm, S.W., 2012. Comparative metagenomics of toxic freshwater cyanobacteria bloom communities on two continents. PLoS One 7, e44002. https://doi.org/10.1371/journal.pone.0044002

Su, X., Steinman, A.D., Tang, X., Xue, Q., Zhao, Y., Xie, L., 2017. Response of bacterial communities to cyanobacterial harmful algal blooms in Lake Taihu, China. Harmful Algae 68, 168–177. https://doi.org/10.1016/j.hal.2017.08.007

Sunda, W.G., Graneli, E., Gobler, C.J., 2006. Positive Feedback and the Development and Persistence of Ecosystem Disruptive Algal Blooms1. Journal of Phycology 42, 963–974. https://doi.org/10.1111/j.1529-8817.2006.00261.x

Suttle, C.A., Harrison, P.J., 1986. Phosphate Uptake Rates of Phytoplankton Assemblages Grown at Different Dilution Rates in Semicontinuous Culture. Can. J. Fish. Aquat. Sci. 43, 1474–1481. https://doi.org/10.1139/f86-184

Teeling, H., Fuchs, B.M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C.M., Kassabgy, M., Huang, S., Mann, A.J., Waldmann, J., Weber, M., Klindworth, A., Otto, A., Lange, J., Bernhardt, J., Reinsch, C., Hecker, M., Peplies, J., Bockelmann, F.D., Callies, U., Gerdts, G., Wichels, A., Wiltshire, K.H., Glöckner, F.O., Schweder, T., Amann, R., 2012. Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton Bloom. Science 336, 608–611. https://doi.org/10.1126/science.1218344

Tromas, N., Fortin, N., Bedrani, L., Terrat, Y., Cardoso, P., Bird, D., Greer, C.W., Shapiro, B.J., 2017. Characterising and predicting cyanobacterial blooms in an 8-year amplicon sequencing time course. ISME J 11, 1746–1763. https://doi.org/10.1038/ismej.2017.58

Walters, W., Hyde, E.R., Berg-Lyons, D., Ackermann, G., Humphrey, G., Parada, A., Gilbert, J.A., Jansson, J.K., Caporaso, J.G., Fuhrman, J.A., Apprill, A., Knight, R., 2015. Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys. mSystems 1, e00009-15. https://doi.org/10.1128/mSystems.00009-15

Wang, J., Zhang, Z., 2020. Phytoplankton, dissolved oxygen and nutrient patterns along a eutrophic river-estuary continuum: Observation and modeling. Journal of Environmental Management 261, 110233. https://doi.org/10.1016/j.jenvman.2020.110233

Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J.R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E.R., Knight, R., 2017. Normalization

and microbial differential abundance strategies depend upon data characteristics. Microbiome 5. https://doi.org/10.1186/s40168-017-0237-y

Willis, A., Woodhouse, J.N., Neilan, B.A., Burford, M.A., 2022. Chapter 12 - Comparative genomics for understanding intraspecific diversity: a case study of the cyanobacterium Raphidiopsis raciborskii, in: Clementson, L.A., Eriksen, R.S., Willis, A. (Eds.), Advances in Phytoplankton Ecology. Elsevier, pp. 415–434. https://doi.org/10.1016/B978-0-12-822861-6.00021-2

Willis, A., Woodhouse, J.N., Ongley, S.E., Jex, A.R., Burford, M.A., Neilan, B.A., 2018. Genome variation in nine co-occurring toxic Cylindrospermopsis raciborskii strains. Harmful Algae 73, 157–166. https://doi.org/10.1016/j.hal.2018.03.001

Woodhouse, J.N., Kinsela, A.S., Collins, R.N., Bowling, L.C., Honeyman, G.L., Holliday, J.K., Neilan, B.A., 2016. Microbial communities reflect temporal changes in cyanobacterial composition in a shallow ephemeral freshwater lake. ISME J 10, 1337–1351. https://doi.org/10.1038/ismej.2015.218

Xie, M., Ren, M., Yang, C., Yi, H., Li, Z., Li, T., Zhao, J., 2016. Metagenomic analysis reveals symbiotic relationship among bacteria in *Microcystis*-dominated community. Front. Microbiol. 7. https://doi.org/10.3389/fmicb.2016.00056

Ye, L., Shi, X., Wu, X., Kong, F., 2012. Nitrate limitation and accumulation of dissolved organic carbon during a spring-summer cyanobacterial bloom in Lake Taihu (China). Journal of Limnology 71. https://doi.org/10.4081/jlimnol.2012.e6

TABLES

Table 1. Summary of environmental variables from Lake Thunderbird, OK, from each sequenced sample. Sonde collected variables were averaged across the photic zone. Variables from water samples are from a depth integrated sample of the photic zone. Numbers are averages with the range in parentheses.

	2016		2017		2018	
	Sailboat	**Fisherman's Point**	**Sailboat**	**Fisherman's Point**	**Sailboat**	**Fisherman's Point**
Total Nitrogen (ug/L)	627.3 (172-833)	665.8 (261-978)	910.2 (322-1430)	950.6 (733-1390)	664.29 (336-913)	660.14 (325-924)
Total Phosphorus (ug/L)	26.79 (21.7-34.1)	25.4 (18.9-37.3)	35.89 (28.6-41.3)	34.6 (24.3-45.8)	45.08 (36.2-68.3)	45.9 (32.6-71)
Chlorophyll-a (extracted ug/L)	23.68 (8.6-32.3)	21.6 (7.8-33.2)	30.34 (14.3-44.3)	28.86 (11.5-43.2)	19.66 (9.59-26.2)	20.78 (11.2-32.04)
Phycocyanin (sonde ug/L)	78.22 (20-114.9)	68.5 (16.9-91.6)	101.4 (49.1-164.9)	94.06 (41.4-166.1)	56.07 (29.7-79.5)	55.46 (29.7-88.8)
turbidity (NTU)	8.88 (7.9-9.8)	9.78 (8.8-11)	7.9 (6.6-11)	8.92 (7.22-11.4)	7.08 (5.4-8.84)	8.04 (6.39-11.2)
water temperature (°C)	29.88 (28.3-31.1)	30.3 (29.3-31.5)	27.9 (26.3-29.3)	28.3 (27.5-29.8)	27.71 (24.5-30.3)	28.13 (24.9-30.7)
dissolved oxygen (mg/L)	7.68 (7.2-8.7)	8.5 (7.8-9.1)	6.98 (4.8-8.4)	7.36 (5.4-8.6)	6.46 (3.9-7.9)	7.11 (5.5-8.5)
pH	8.64 (8.5-8.8)	8.7 (8.6-8.8)	7.96 (7.6-8)	7.96 (7.8-8.2)	7.74 (7.4-7.9)	7.77 (7.6-8)
max wind speed (m/s)	2.14 (1.4-3.4)	2.44 (0.8-4.3)	3.24 (1.7-4.6)	4.14 (2.6-6.1)	3.11 (0-3.2)	4.61 (0.7-6.8)
average wind speed (m/s)	1.24 (0.7-1.7)	1.6 (0.3-3.3)	1.86 (1.1-3.1)	2.44 (0.8-4.3)	1.79 (0-3.2)	3.36 (0.6-5.8)
air temperature (°C)	31.38 (30-36)	30.8 (28.2-32.1)	29.86 (27.8-33)	30.58 (28.5-32.2)	29.34 (20.4-33.8)	29.94 (27-34)
total depth (m)	2.29 (2-2.5)	2.2 (2-2.3)	2.08 (1.9-2.2)	1.96 (1.8-22)	2.24 (2.05-2.5)	1.96 (1.4-2.5)

Table 2. Environmental variable p-values associated with the bacteria community NMDS environmental vectors. Significant p-values are bolded.

Environmental Variable	NMDS axses 1 and 2 p-value	NMDS axses 1 and 3 p-value
max wind speed (m/s)	0.227	**0.002**
average wind speed (m/s)	0.351	**0.004**
air temperature °C	0.082	**0.046**
turbidity (NTU)	0.247	**0.038**
total nitrogen (ug/L)	**0.001**	**0.002**
total phosphorus (ug/L)	**0.011**	**0.001**
water temperature °C	0.287	**0.001**
dissolved oxygen (mg/L)	0.114	**0.001**
ph	0.106	**0.001**
chlorophyll-a (ug/L)	**0.002**	**0.001**
phycocyanin (ug/L)	**0.006**	**0.001**
Nitrogen:Phosphorus (N:P)	**0.007**	0.054
total depth (m)	0.529	0.777

FIGURES

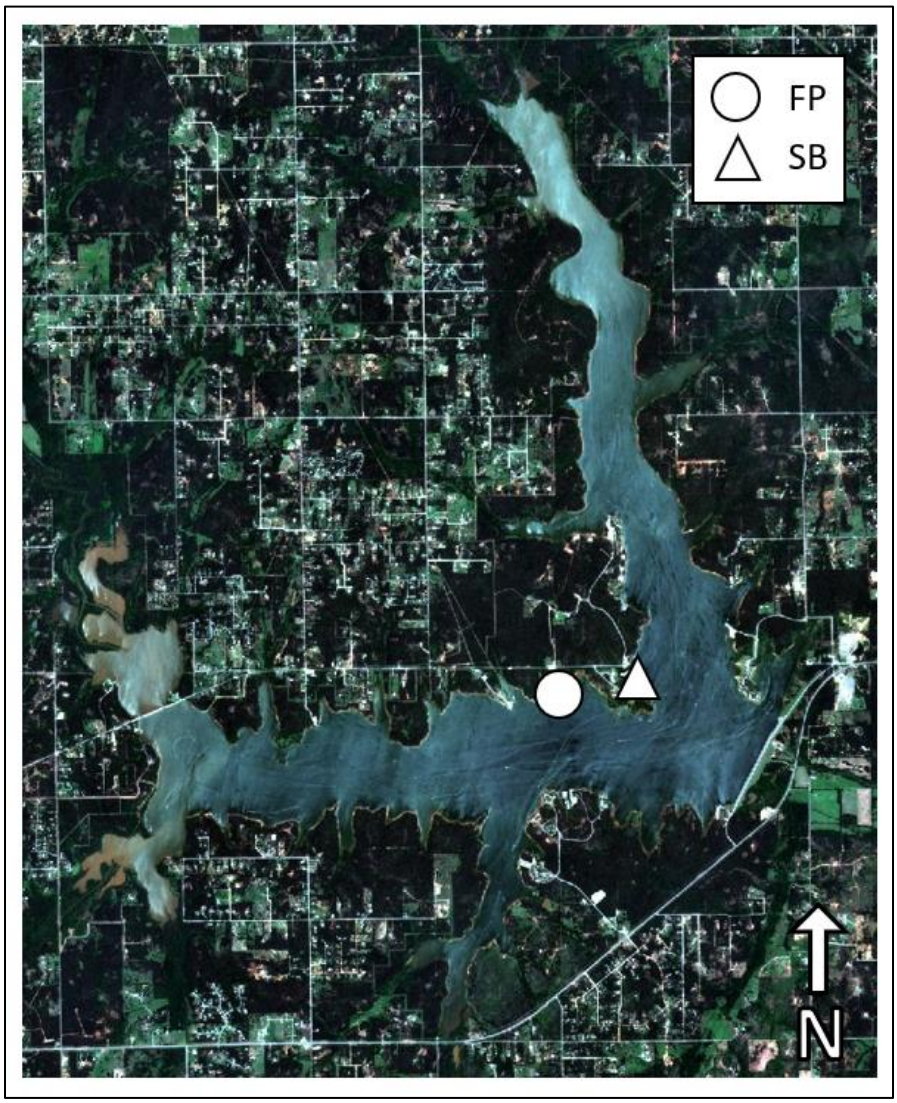


Figure 1. Sites sampled on Lake Thunderbird, OK, USA. Fisherman's Point (FP, circle; 35.23, -97.25) was sampled from a fishing pier that extends approximately 40m into the lake. Sailboat (SB, square; 35.23, -97.24) was sampled from a floating dock next to the boat ramp that extends ~ 43 m.

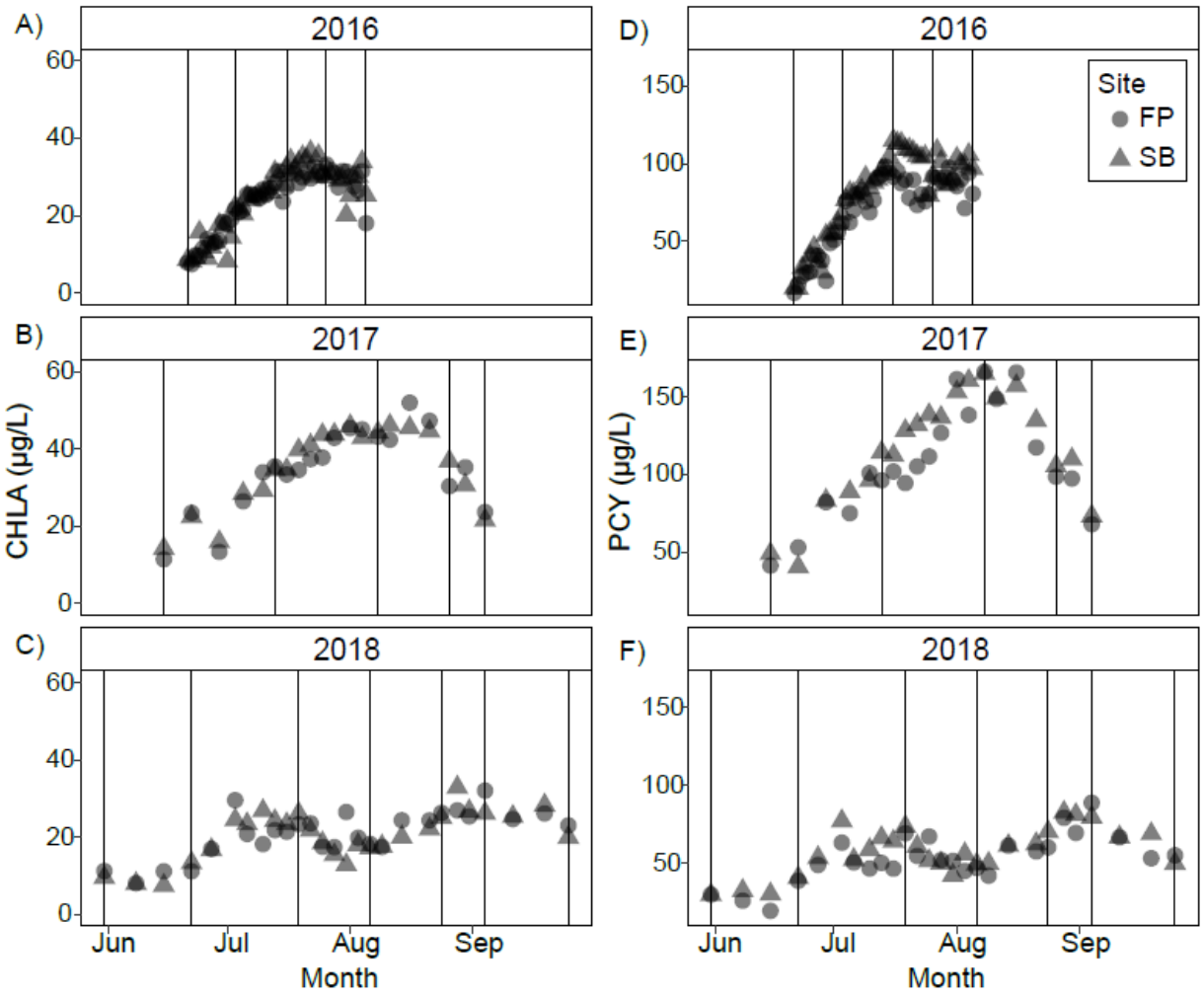


Figure 2. Pigment concentrations across three years of sampling: 2016 chlorophyll-a (CHLA) (A), 2017 CHLA (B), 2018 CHLA (C), 2016 phycocyanin (PCY) (D), 2017 PCY (E), and 2018 PCY (F). Phycocyanin (PCY) was measured using a Hydrolab sonde and chlorophyll-*a* (CHLA) was extracted. The sites are indicated by different shapes, circle for Fisherman's Point and triangle for Sailboat. Darker areas indicate increased overlap of points. Vertical lines indicate bloom phase samples used for sequencing and subsequent data analysis.

Figure 3. Relative abundance of only cyanobacteria genera over the course of three years of sampling. The first column is Fisherman's Point (FP) and the second column is Sailboat (SB). The first row is year 2016, the second row is year 2017, and the third row is year 2018. Displayed are the top 9 most dominate cyanobacteria (all had abundance >50). Bar position corresponds with sampling date.

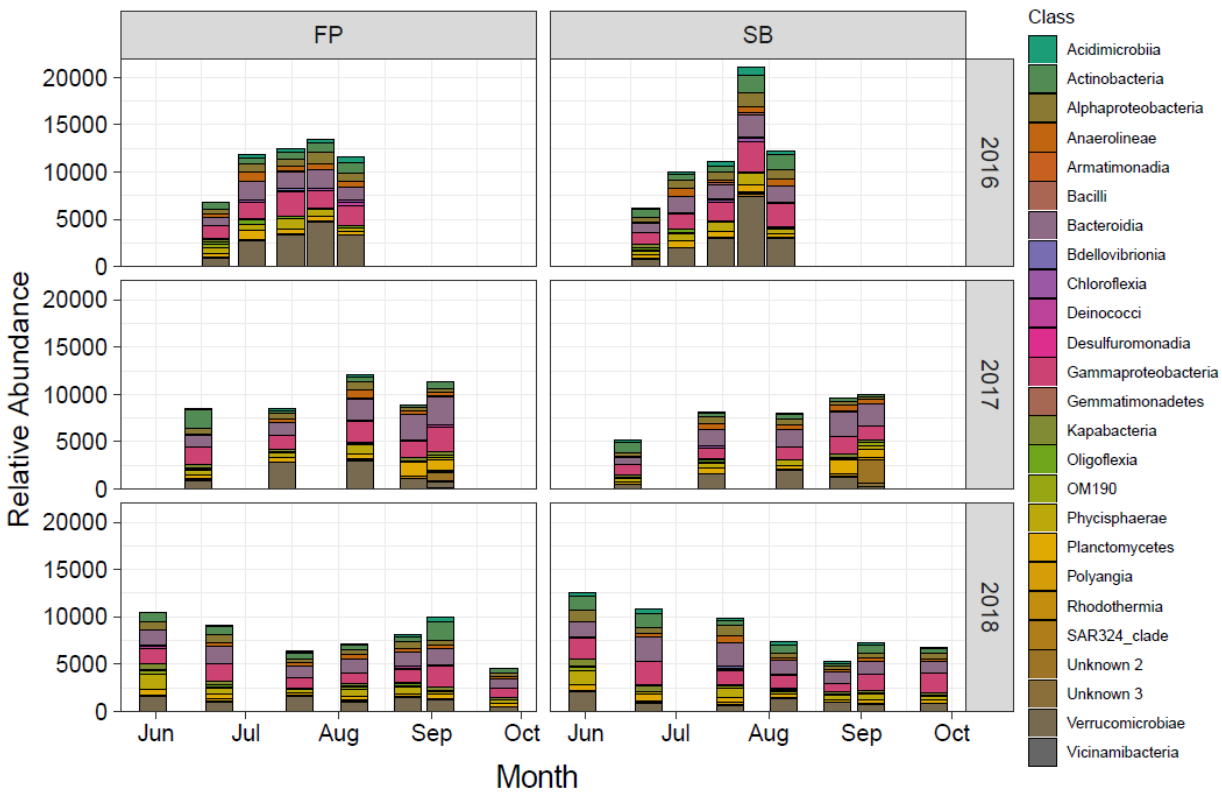


Figure 4. Relative abundance of only bacteria classes over the course of three years of sampling. The first column is Fisherman's Point (FP) and the second column is Sailboat (SB). The first row is year 2016, the second row is year 2017, and the third row is year 2018. Displayed are bacteria with an abundance of at least 100 in one sample. Bar position corresponds with sampling date.

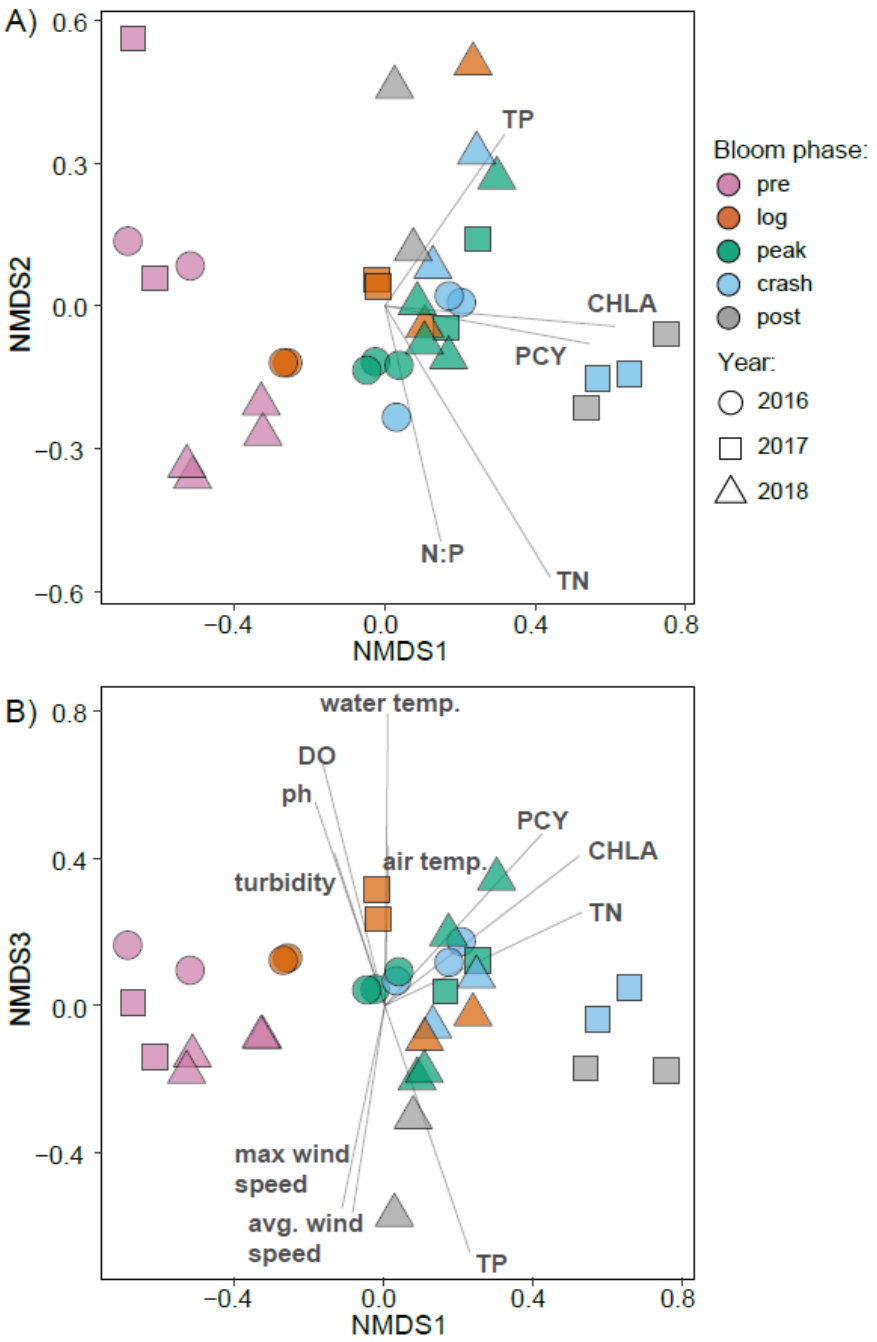


Figure 5. NMDS ordination (Bray-Curtis, stress = 0.07, k = 3) examining relationship between the bacteria community composition (no cyanobacteria) and environmental variables (A - NMDS axes 1 and 2; B - NMDS axes 1 and 3). Each point represents a unique sample, and these communities are colored by pre-determined bloom phase and year for visual comparison. The distance between points represents the amount of similarity (or dissimilarity) in bacteria

community composition. PERMANOVA indicated peak bloom communities were significantly different from pre bloom communities ($p = 0.001$). Significant environmental variables are displayed as vectors (Table 2), where the strength of the relationship is represented by length of the vector and the end with the label indicates the highest values of the gradient.

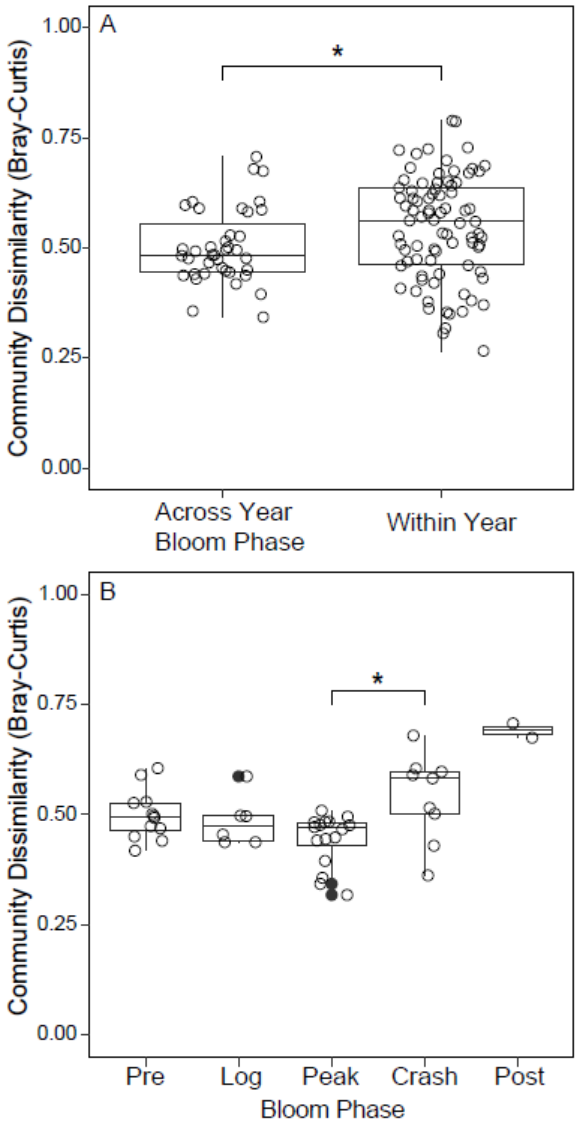


Figure 6. (A) Results of the comparison of community dissimilarity between bloom phase and year. The points for bloom phase represent all possible comparisons of a given community with similar bloom phases across year, controlling for year (ex., all pairwise comparisons of pre-bloom phase across all three years, except pre-bloom samples that occurred in the same year). Year points represent all possible comparisons of communities within the same sampling year (three unique sampling years, 2016, 2017, and 2018). Bacteria community dissimilarity was significantly different between bloom phase and year (Welch's t-test: t = -2.3, df = 101.4, p-value = 0.02), with the across year group having greater dissimilarity (high Bray-Curtis values).

(B) Bacteria community dissimilarity for each unique bloom phase. Excluding the post samples due to the extreme difference in sample size, the other bloom phases bacteria community dissimilarity were compared and found to be significantly different by phase (ANOVA, df =3, SumSq = 0.06, MeanSq = 0.02, F = 4.2, p = 0.01). The peak communities were significantly more similar to each other than the crash communities where to other crash communities (TukeyHSD p = 0.007).

Chapter 3 - Ground-based remote sensing provides alternative to satellites for monitoring cyanobacteria in small lakes

Formatted for and in review at *Water Research*

Katherine V. Cook,[a,b], Jessica E. Beyer,[a,b], Xiangming Xiao,[d], and K. David Hambright,[a,b,c]

[a]Plankton Ecology and Limnology Laboratory, Department of Biology, University of Oklahoma, Norman, USA;

[b]Program in Ecology and Evolutionary Biology, Department of Biology, University of Oklahoma, Norman, USA;

[c]Geographical Ecology, Department of Biology, University of Oklahoma, Norman, USA;

[d]Center for Earth Observation and Modeling, Department of Microbiology and Plant Biology, University of Oklahoma, Norman, USA

KEYWORDS: Cyanobacteria, Remote sensing, Chlorophyll-*a*, Phycocyanin, Water Quality, Harmful algal bloom, Monitoring

ABSTRACT

Cyanobacteria are the most prevalent bloom-forming harmful algae in freshwater systems around the world. Adequate sampling of affected systems is limited spatially, temporally, and fiscally. Remote sensing using space- or ground-based systems in large water bodies at spatial and temporal scales that are cost-prohibitive to standard water quality monitoring has proven to be useful in detecting and quantifying cyanobacterial harmful algal blooms. This study aimed to identify a universal multispectral reflectance model that could be used for rapid, remote detection and quantification of cyanoHABs in small- to medium-sized productive reservoirs, such as those typical of Oklahoma, USA. We used Landsat satellite reflectance and *in-situ* pigment data spanning 16 years from 38 reservoirs in Oklahoma to construct empirical linear models for predicting concentrations of chlorophyll-*a* and phycocyanin, two key algal pigments commonly used for assessing total and cyanobacterial algal abundances, respectively. We also used ground-based hyperspectral reflectance and *in-situ* pigment data from seven reservoirs across five years in Oklahoma to build multispectral models predicting algal pigments from newly defined reflectance bands. Our Oklahoma-derived Landsat- and ground-based models outperformed established reflectance-pigment models on Oklahoma reservoirs. Importantly, our results demonstrate that ground-based sensors and multispectral models were far superior to Landsat-based models for monitoring cyanoHABs in highly productive, small- to mid-sized reservoirs in Oklahoma. While satellite-based remote sensing approaches have proven effective for relatively large systems, our results indicate that ground-based remote sensing may offer better cyanoHAB monitoring for small or highly dendritic turbid lakes, such as those throughout the southern Great Plains, and thus prove beneficial to efforts aimed at minimizing public health risks associated with cyanoHABs in supply and recreational waters.

1. INTRODUCTION

Cyanobacterial harmful algal blooms (cyanoHABs) in freshwaters are a topic of global concern (Paerl and Barnard, 2020). Blooms are increasing in frequency and magnitude due, in part, to increased anthropogenic nutrient loading (Paerl and Paul, 2012). CyanoHABs are particularly problematic because many species of harmful algae produce toxins that pose health risks to humans and other animals through exposure by toxin consumption, inhalation, and topical contact (Hambright et al., 2014; Hilborn and Beasley, 2015; Plaas and Paerl, 2021). While some cyanotoxins may be only dermatotoxic, producing mild to moderate allergic reactions, common cyanotoxins, such as microcystin, cylindrospermopsin, saxitoxin, and anatoxin are powerful hepato- and neurotoxins that can cause chronic illness and death. With increasing threats of cyanoHAB development in inland waters, there is an equally increasing need for rigorous monitoring for cyanobacteria and their toxins to minimize public health risk associated with cyanoHABs (Paerl and Barnard, 2020).

Unfortunately, cyanoHAB monitoring today is highly insufficient for the needs of health risk management (Almuhtaram et al., 2021). Historically, cyanoHAB assessment has been based on standard limnological monitoring of water quality with focus on criteria, such as water clarity or concentrations of chlorophyll-*a* and nutrients. Monitoring of cyanoHAB densities and cyanotoxins requires specialized skill and instrumentation and is thus generally conducted routinely only in lakes that provide water supply to large metropolitan communities. Similar routine cyanoHAB monitoring for the vast majority of lakes and reservoirs in rural settings is physically and economically impractical (Almuhtaram et al., 2021). High-frequency autonomous monitoring is possible with expensive, high-maintenance monitoring platforms (Coad et al., 2014). However, sufficient spatial coverage, particularly for moderate- to large-sized dendritic reservoirs, with tens to hundreds of semi-enclosed bays and coves (areas with high recreational

use), would be prohibitively expensive. To minimize potential health risks associated with cyanoHAB toxin exposure in such systems, lake management and public health agencies need new tools that are amenable to broad and simple implementation across multiple and diverse systems, particularly those supporting high recreational tourism.

An early solution to the monitoring needs of public health protection emerged as remote measurement of solar reflectance by satellite-based sensors was shown to be effective and economically beneficial in assessing general chlorophyll-*a* concentrations in surface waters (Gons et al., 2002; Stroming et al., 2020). Satellite imagers have the potential to allow monitoring without the expensive person-hours and equipment required for a physical visit to a lake or a permanent autosampler/profiler, due to their large spatial coverage (~900-90,000 m^2) which can collect images for many lakes simultaneously and their fine temporal resolution (daily to fortnightly flyovers, depending on satellite). Obviously, satellites are expensive, but many currently in use for earth observations were launched by public entities that offer users open-access data (images), thus providing lake managers a low-cost option for surface water observation.

Successful application of satellite-based remote sensing to cyanoHAB assessment, particularly for large waterbodies (Coffer et al., 2020; Urquhart et al., 2017; Wynne et al., 2010), has fueled a search for universal models can predict concentrations of chlorophyll-*a* and phycocyanin, two key algal pigments commonly used for assessing total and cyanobacterial algal abundances (Shi et al., 2019). However, there are well-known limitations that may interfere with a satellite-based approach, including significant loss of usable images due to cloud cover (Ju and Roy, 2008) and sun-glint contamination (Overstreet and Legleiter, 2017). Moreover, satellite sensors frequently used for remote sensing of HABs (MODerate resolution Imaging

Spectroradiometer, MODIS, aboard the National Aeronautics and Space Administration's (NASA) Terra and Aqua satellites; MEdium Resolution Imaging Spectrometer, MERIS, aboard the European Space Agency's (ESA) Envisat-1 satellite; and Ocean Land Colour Instrument, OLCI, a follow-up to ESA's MERIS) were built for a combination of atmosphere, terrestrial, and oceanic applications (Barnes et al., 1998; Nieke et al., 2012; Rast et al., 1999). The sensors on these satellites quantify reflectance from relatively large areas (e.g., 250 × 250 to 500 × 500 m) in each pixel. This resolution is not amenable for observing most inland lakes and reservoirs that are either too small or are highly dendritic in structure, such that many water pixels can be contaminated by littoral, shoreline, and terrestrial reflectance (Verpoorter et al., 2014; Wetzel, 2001). Additionally, water levels in smaller water bodies fluctuate seasonally, making the delineation between water and non-water contamination notoriously difficult (Zou et al., 2017), especially with low spatial resolution instruments. These spatial resolution limitations have led to a lack of data for smaller waterbodies in the remote sensing literature and the call for development and use of satellites with spatial resolutions of 30 × 30 m or smaller (Beck et al., 2016; Urquhart et al., 2017; Coffer et al., 2020).

An alternative to satellite-based remote sensing is the use of ground-based sensors, which do not suffer from the previously mentioned satellite limitations, like atmospheric interference (Wu et al., 2019). Ground-based instruments with multispectral sensors have many advantages to satellites, like the ability to capture images with very fine spatial resolution, flexible temporal resolution, and rapid data turn-around time; studies suggest ground-based instruments will provide better capability for monitoring cyanoHABs than satellite-based remote sensing (Wu et al., 2019). There are commercially available sensors for use on ground-based instruments, but most were designed for terrestrial application. No commercially available sensors have a band

designed specifically for phycocyanin. Fernandez-Figueroa et al. (2021) found commercially available cameras and multispectral sensors were more sensitive to chlorophyll-*a* than phycocyanin in eutrophic ponds. Multispectral sensors built for detection of phycocyanin are needed for adequate monitoring of cyanoHABs (Almuhtaram et al., 2021; Fernandez-Figueroa et al., 2021).

In this study, we sought a universal multispectral reflectance model that could be used for rapid, remote detection and quantification of cyanoHABs in reservoirs that are relatively small or highly dendritic, productive, and often turbid, like those in Oklahoma. We aimed to explore the effectiveness of satellite-based multispectral models for predicting chlorophyll-*a* and phycocyanin concentrations in Oklahoman reservoirs in comparison to new ground-based multispectral models derived from the same Oklahoma systems. Our results show that the published models had little to no predictive power with respect to assessing cyanoHABs in a study system outside those used to establish the models, and that ground-based sensors and multispectral models, particularly for phycocyanin, were far superior to Landsat-based sensors and models for assessing cyanoHABs in small, dendritic, productive reservoirs in Oklahoma.

2. METHODS

2.1 Study System

Oklahoma is an ideal study system for turbid, small- to medium-sized reservoirs. The majority of Oklahoma lakes are smaller than the average American football field (i.e., ca. 0.5 ha, Zou et al., 2017) or are highly dendritic, with high shoreline to surface area ratios (aka, shoreline development; (Wetzel, 2001)). Oklahoma lakes are typical of lakes in the southern Great Plains with high turbidity, high chlorophyll-*a*, and frequent cyanoHABs (Oklahoma Water Resources

Board, 2017). Lakes in Oklahoma are used primarily for water supply and recreation, and therefore monitoring harmful algal blooms is necessary to protect public health (Smithee et al., 2012).

With these characteristics in mind, we compared reflectance-algal pigment models using Landsat satellite data and ground-based hyperspectral reflectance. We built empirical models relating satellite spectral reflectance to chlorophyll-*a* (Table. S1) and phycocyanin (Table. S2) concentrations from 38 lakes (N = 1060 paired observations) sampled between 2001 – 2017 and ground-based reflectance to chlorophyll-*a* and phycocyanin concentrations from 7 lakes (N = 127) sampled between 2012 – 2016 (see Fig. 1A and B for map of lake locations). We also compared our Oklahoma-specific models to published models.

2.2 Satellite reflectance data

The Enhanced Thematic Mapper Plus (ETM+) and Operational Land Imager (OLI) are NASA multispectral sensors onboard the Landsat 7 and Landsat 8 satellites, respectively, that consist of different bands (ETM+ has eight bands and OLI has 11 bands) that record select areas of the electromagnetic spectrum (please see Irons et al. (2012) for details on the sensors and bands). These two satellites were chosen over other potential sensors (e.g., MERIS, OLCI) for this study due to the relatively small pixel size (30 × 30m) of the ETM+ and OLI sensors and the eight-day revisit time between the satellites. Sentinel-3A OLCI was not chosen due to its large pixel size of 300 m which is difficult to resolve on small, dendritic Oklahoma reservoirs (see Fig. S6 for Landsat and Sentinel-3A image comparison) and it only became operational in 2016, years after much of our in-situ water sampling.

Surface reflectance data were obtained from EROS Data Center, Sioux Falls, SD (USGS) for each site-date combination. Surface reflectance data are generated by the USGS from Level 1 Digital Number (DN) products by applying a MODIS/6S atmospheric correction routine using the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) software (Masek et al., 2006). Images (N = 247) were matched with *in-situ* measurements within a ±7-day window of the sampling date, ensuring a difference of no more than seven days between the satellite image and the *in-situ* measurements. This window was informed by a concurrent study measuring temporal autocorrelation in cyanobacterial blooms in Oklahoma (Beyer and Hambright, 2017) which showed that proxies (chlorophyll-*a* and phycocyanin) for cyanobacterial abundance during bloom periods were strongly temporally autocorrelated over a period of 10-11 days.

Images were visually screened for cloud cover and excluded if the lake area had >10% cloud cover. Since July 14, 2003, the Landsat 7 ETM+ Scan Line Corrector has been turned off due to an instrument failure, leaving stripes of missing data across images (Markham et al., 2004). To minimize the potential impact of missing data, and to account for the anchored swing radius of the boat under changing winds and currents, we averaged reflectance values from nine pixels (a 3 × 3 grid consisting of the sampling point pixel and the eight surrounding pixels) for each site. This approach maximized the number of usable data points and achieved the most representative sample for each reflectance measure (Baban, 1993). Reflectance data were extracted from images using R (R Development Core Team, 2014) (see Supplemental for code). The data were then plotted and screened for high reflectance outliers (reflectance greater than 1). Water reflectance is much lower than land reflectance, so high reflectance values could indicate shoreline, or bottom interference, or an object (e.g., a boat) in the pixel.

2.3 Ground-based reflectance data

We collected *in-situ* hyperspectral reflectance data from seven lakes (see Fig. 1C for map) over a period spanning 2012-2016 using an ASD FieldSpec spectroradiometer (Malvern Panalytical, Malvern, United Kingdom). At each lake, the ASD was calibrated using a Spectralon reflectance panel (Labsphere, North Sutton, NH, USA). At each pelagic site (n = 14) and shoreline site on Lake Thunderbird (n = 12), the optical probe was held over the non-shaded side of the boat or boat dock at a 45-degree angle to the water and five spectra were recorded. These five spectra were averaged for each site to minimize the impact of sun glint or waves and the resulting spectra were used in downstream analyses. For many of these lakes, we had repeat visits to multiple sites (Table. S3). In total, we collected hyperspectral data from 132 site-date combinations.

Quality control and quality assurance were performed on the spectra. Spectra were visually examined, and six spectra were discarded: three due to the loss of the raw files, one due to an abnormal horizontal spike at 1000 nm, one due to abnormal low reflectance values, and another due to an outlier in the limnological data. Each spectrum was matched with same day *in-situ* pigment data. One additional site-date combination was discarded due to lack of corresponding limnological data. The final working sample size was 127 independent hyperspectral measurements.

2.4 *In-situ* surface water pigment data

Each satellite image and ground-based measurement has a corresponding pigment measurement from an Oklahoma lake. These data were collected by personnel in the Plankton Ecology and Limnology Lab (PELL) at the University of Oklahoma, as well as the Oklahoma

Water Resources Board (OWRB) and the Grand River Dam Authority (GRDA). For the satellite dataset, 38 lakes were sampled between 2001 – 2017 for a total of 1060 chlorophyll-*a* data points (Fig. S1; Table. S1) and 97 phycocyanin data points (Fig. S3; Table. 2). For the ground-based measurements, seven lakes were sampled between 2012 – 2016 for a total of 127 chlorophyll-*a* (Fig. S4; Table. S3) and 127 phycocyanin data points (Fig. S5; Table. S3).

Lake sampling by PELL is briefly summarized here, for details see Hambright et al. (2015). At each site, a Hydrolab DS5X (OTT Hydromet, Kempten, Germany) sonde was used to measure phycocyanin (PCY) concentrations (starting in 2009). Although PCY in nature will generally be bound within cells, the Hydrolab PCY sensor was calibrated using serial dilutions of biologically relevant levels of pure PCY solution (Sigma Aldrich) and thus provides an underestimate of the actual amount of PCY in lakes. Nevertheless, relative changes can still provide meaningful insight into seasonality and site-to-site comparisons. Such PCY measurements have previously been found to correlate with cyanobacterial biomass (Thomson-Laing et al., 2020). At deeper sites, measurements were taken in the upper 10 m of the water column at 1-m intervals, and at shallow sites measurements were taken at 0.5-meter intervals through the entire water column (on average around 2m). Phycocyanin measurements were averaged across the photic zone (as determined by a LI-COR light meter (LI-COR Biosciences, Lincoln, NE, USA) or 2.5 times the Secchi depth). For acetone-extracted chlorophyll-*a*, depth-integrated water samples were taken from the photic zone depth or the entire water column, whichever was shallower, and placed into acid-washed, deionized- and sample-rinsed Nalgene bottles, and kept on ice prior to filtration. The water was filtered onto 25-mm-diameter GF/F filters and stored in the freezer until chlorophyll-*a* extraction. In a reduced light environment, the filters were homogenized with a small volume of 70% acetone using a mortar and pestle, then

incubated in 70% acetone at 4-8˚C for 6-24 hours. Samples were centrifuged prior to analysis on a Turner Designs 700 (pre-2012) or Trilogy (2012-present) bench-top fluorometers (EPA, 1997, section 4.2).

The OWRB and GRDA collected water samples from the upper 0.5 m for chlorophyll-*a* analysis. The agencies did not measure phycocyanin and both used the previously described EPA chlorophyll-*a* acetone extraction method. For the satellite dataset, OWRB contributed 836 data points, GRDA contributed 56, and PELL contributed a total of 168. For the ground-based dataset, all 127 chlorophyll-*a* and phycocyanin pigment data points were contributed by PELL.

2.5 Landsat multispectral models

To quantify cyanoHABs in Oklahoma, we built empirical linear regression models using the chlorophyll-*a*, phycocyanin, and satellite reflectance values. We used Landsat band ratios as our predictors because spectral ratios have been shown to be more robust than single bands (Vincent et al., 2004). We eliminated potential collinearity in the models using variance inflation factors (VIF) by sequentially removing the predictors with the highest VIF score until all scores were at the predetermined threshold of 10 (Zuur et al., 2010). The spectral ratios used to determine the final chlorophyll-*a* model included the following band ratios: Blue:Green, Red:Blue, Green:Red, Blue:Near Infared (NIR), and NIR:Red; and in the PCY model: Blue:Green, Green:Red, NIR:Red, and Red:NIR. We used Bayesian Information Criterion (BIC) to select the model that explained the most variance in the response variable using the fewest predictors (Schwarz, 1978). All statistical analyses were completed in the R environment (version 4.2.1).

To test how published models effectively predicted cyanoHABs in Oklahoma lakes, we applied a set of chlorophyll-*a* algorithms adapted to Landsat 7 from the Landsat 8 models (Landsat 8 bands 2–6 match Landsat 7 bands 1–5) in Beck et al. (2016) (Table. 1). The algorithms from the literature include the Normalized Difference Chlorophyll Index (NDCI) (Mishra and Mishra, 2012), the Surface Algal Bloom Index (SABI) (Alawadi, 2010), the Fluorescence Line Height algorithm focusing on the blue band (FLH blue) (Zhao et al., 2010), the two-band algorithm (2BDA) (Gitelson et al., 2003), the three-band algorithm (3BDA) (Dall'Olmo and Gitelson, 2005) and the three band-like algorithm (KIVU) (Brivio et al., 2001; Kneubühler et al., 2007). We also used one established phycocyanin Landsat algorithm from the literature (Vincent et al., 2004) to compare with our Landsat phycocyanin model (Table 1). These algorithms were chosen in part because they were built using Landsat bands or because they have been shown to successfully predict chlorophyll-*a* or phycocyanin.

Each of the previously mentioned chlorophyll-*a* models was applied to our Oklahoma satellite reflectance data set (N = 1060) and compared to *in-situ* chlorophyll-*a* values to determine model accuracy. As per the original use, each algorithm was solved prior to regressing (meaning there is a single regression coefficient for each algorithm). For the Vincent et al. (2004) phycocyanin algorithm, we fit a regression coefficient to each spectral ratio as was done in the original model. To assess the fit of the regressions, we compared the *in-situ* observed pigment values with the predicted algorithm pigment values and evaluated the relationship using the root mean square error (RMSE; lower values indicate better model fit) and Pearson's r correlation test (range from +1 to –1 with numbers closer to 0 indicating no correlation) (Table 3). We also report the adjusted R^2 of each regression as this is a direct measure of model predictability (ranges from 0-1 with higher values being more predictive).

In addition to the above described empirical models, we built random forest models to explore the utility of machine learning as an approach for generating useful predictive models. We used the randomForest function in the randomForest package in R (Liaw and Wiener, 2002) to grow random forests with ntree = 500. For both chlorophyll-*a* and phycocyanin models we used the non-collinear predictors described above for the linear models. We chose regression random forest because we are most interested in prediction across a range of pigment values and they are more comparable to previously described linear models. We evaluated model performance similarly to the linear models using RMSE and R^2 but in addition, we tested significance against a null model. We extracted the performance statistics and significance using the rfUtilities package in R using the rf.regression and rf.significance functions, respectively (https://cran.r-project.org/web/packages/utility/index.html).

2.6 Ground-based multispectral methods

We built a semi-empirical model using the ground-based measurements by first selecting five bands based on spectral properties of optical constituents of cyanobacteria and other non-target components that could affect reflectance of light from Oklahoma reservoirs, specifically, chlorophyll-*a*, phycocyanin, and turbidity. The medians of the bands were based on published values (Matthews, 2011) and the width of the bands were optimized by comparing correlations of candidate bands with our measured optical constituents (Table. 2). We calculated the average reflectance over the wavelengths included in the band. Using the bands (described in Table 2), we constructed candidate linear models for chlorophyll-*a* including all possible combinations of ratios of the five bands and selected the best model using the corrected Akaike information criterion (AICc). As prediction was our goal, and not inferring causality, we moved forward in

analyzing the model with the lowest AICc value. This process was repeated for phycocyanin. We compared our models (using adjusted R^2 and RMSE) with the chlorophyll-*a* and phycocyanin nested-band models from Randolph et al. (2008) which have been found to accurately predict chlorophyll-*a* and phycocyanin concentrations using hyperspectral data (see Table. 1 for algorithms and see eq. 4 and 5 in Randolph et al. (2008) for complete algorithms). We also compared our models with the Cyanobacterial Index (CI) and updated Cyanobacterial Index (CI-multi; Matthews et al., 2012) as defined in Coffer et al. (2020), which allows chlorophyll-*a* reflectance to be attributed to cyanobacteria based off the spectral shape at wavelength 665. To implement the CI and CI-multi algorithms, we first resampled our ground-based hyperspectral data to match the bands of MERIS by calculating the average reflectance over the wavelengths included in each band. Then we calculated CI and CI-multi following the methods of Coffer et al. (2020). CI and CI-multi were converted to chlorophyll-*a* (µg/L) using the equation in Tomlinson et al. (2016).

3. RESULTS

3.1 Satellite models

The best chlorophyll-*a* model built using Oklahoma data (N = 1060) contained four Landsat ETM+ and OLI band ratios: Blue:Green, Red:Blue, Blue:NIR, and NIR:Red Table 3.7).

$$\text{Chl-}a = 71.8 - 37.3(\text{Blue:Green}) - 10.1(\text{Red:Blue}) - 6.6(\text{Blue:NIR}) - 3.8(\text{NIR:Red}), \quad \text{(eq. 1)}$$

Based on the adjusted R^2, RMSE, and Pearson's r correlation coefficient, our Oklahoma-derived model outperformed all other models (Table 3.1-6), although they had relatively low predictive

power. Of the literature algorithms, the KIVU algorithm (adapted from Beck et al. 2016) performed the best, having the lowest RMSE and highest adjusted R^2 values, although it had low predictability. The remainder of the literature models performed poorly with little to no predictive power for algal pigments in Oklahoma lakes (Table 3.1-6).

The initial performance of our chlorophyll-*a* model was poor. This poor performance could be due in part to the skewed nature of our data, with the majority of the chlorophyll-*a* values falling between 2-30 µg/L (Fig. S1). To test if this skew in the data was causing poor prediction in the model we resampled chlorophyll-*a* data to create a more even distribution. Specifically, the original data was sorted into 30 intervals. The intervals containing the chlorophyll-*a* values 2-37 µg/L (intervals 2-8) were randomly subsampled without replacement to reduce the whole data set by approximately half (N = 511) (Fig. S2). We retained all higher chlorophyll-*a* values, because while we believe the original data distribution is reflective of total chlorophyll-*a* values seen annually in Oklahoma lakes, we are interested in predicting the higher chlorophyll-*a* values that would be associated with a bloom. We then applied the best chlorophyll-*a* model (Eq. 1) to the new data set (N = 511) and found the predictability did not improve from the full data set. The literature algorithms were also tested on the reduced dataset and there was no improvement in performance (Table S4.1-6). All models, including our own, showed systematic bias where samples with medium to high (30-150 µg/L) observed chlorophyll-*a* values had much lower predicted values, never exceeding 30 µg/L (Fig. 2A).

The best phycocyanin (N = 97) model built using Oklahoma data used two band ratios: Blue:Green and Red:NIR (Table 3.10).

$$PCY = 146.7 - 124.5(Blue{:}Green) - 16.7(Red{:}NIR), \quad \text{(eq. 2)}$$

This phycocyanin model underpredicted at higher values, meaning it predicted much lower phycocyanin values than were observed (Fig. 2B). We also tested a phycocyanin model from the literature, the Vincent et al. (2004) model, on the Oklahoma dataset. The Vincent et al. (2004) model performed better than the model from this paper (Table 3.9).

The chlorophyll-*a* random forest model constructed using satellite data performed poorly, only explaining 3.07% of the variance (Table 3.11) and it was not significantly different when tested against a null model built with the same data. The phycocyanin random forest explained 32.9% of the variance and performed better than our linear two-band ratio model but not as well as Vincent's satellite phycocyanin model (Table 3.9-10, 11).

3.2 Ground-based models

To test the efficacy of ground-based sensors at quantifying cyanoHABs in Oklahoma we built a semi-empirical multispectral model using hyperspectral data taken from Oklahoma lakes over a variety of conditions to predict chlorophyll-*a* and phycocyanin concentrations, respectively (n = 127). The best chlorophyll-*a* model included six band ratios (1:3, 1:5, 2:3, 3:4, 3:5, 4:5; see Table 2 for band information, see Eq. S1 for full equation). The CI, CI-multi, and Randolph et al. (2008) chlorophyll-*a* nested-band-ratio model were also applied to the Oklahoma lakes data and resulted in slightly lower performance compared to our six-band ratio multispectral model (Table 4.1-4; Fig. 3).

Our best phycocyanin model included seven band ratios (1:2, 1:5, 2:3, 2:4, 3:4, 3:5, 4:5; see Table 2 for band information, see Eq. S2 for full equation). Comparison of the predictions of our phycocyanin seven-band ratio multispectral model with those of a common nested-band-ratio

model (Randolph et al., 2008), CI, and CI-multi revealed increased accuracy with the seven-band ratio approach (Table 4.5-8; Fig. 4).

4. DISCUSSION

The frequency and magnitude of cyanoHABs are increasing globally, in pace with climate and land-use change, and increasing nutrient pollution (Huisman et al., 2018). Detecting and tracking blooms in a timely manner for risk management has proven to be difficult and costly (Almuhtaram et al., 2021b). Remote sensing has the potential to alleviate the insufficiencies of traditional monitoring by providing fast and less expensive information while allowing the end user to monitor many lakes simultaneously. Ground-based remote sensing has increased in popularity in recent years and appears more adaptable for many different systems and scenarios, thus it could be better for monitoring cyanoHABs. Here, we assessed the possibility of using satellite- and ground-based remote sensing for quantifying cyanoHABs in Oklahoma reservoirs.

We were unsuccessful in finding a universal satellite pigment model for Oklahoma lakes. Our Landsat-based models for chlorophyll-*a* derived from Oklahoma reservoirs were only marginally better than published algorithms, and none of them were sufficiently predictive ($R^2 \leq 0.035$) for implementation in monitoring programs. Even though we were using mostly Landsat 7 data, our results corroborate those of Beck et al. (2016), who applied these algorithms to Landsat 8 simulated imagery and found poor performance across the board except with the Fluorescence Line Height violet algorithm. We were unable to use this algorithm because Landsat 7 lacks a comparable band to Landsat 8's Coastal/Aerosol Band 1 (0.435-0.451 nm) used in the algorithm. While it was our 'best' performing literature algorithm, we expected better

performance from the KIVU 3-band-like algorithm because it performed well on Lake Garda when adapted from MERIS bands to Landsat 5 TM bands (Brivio et al., 2001), which are comparable to Landsat 7 ETM+ bands. The other algorithms were originally created using other sensors, such as MERIS and MODIS, or hyperspectral data that was translated to match Landsat satellite bands. This cross-sensor translation was likely another contributor to the poor performance of the models (Beck et al., 2016).

The poor prediction of the algorithm may also reflect the size of our data set, consisting of 1,060 paired data points (much larger than previous studies) and a large number of waterbodies used (N = 38). The reflectance values we measured ranged from 0.0025 to 0.37 and Landsat surface reflectance values ranged from 0 to 1, meaning instrument saturation is unlikely the cause of the poor performance (*Landsat surface reflectance data*, 2015). Instead, this prediction bias is likely due to the distribution of the original data, where the majority of the chlorophyll-*a* values ranged 2-30 µg/L. Considering the wide temporal nature of our data (2001-2017; all months retained), we believe this chlorophyll-*a* distribution is representative of natural conditions. Even when the data were subsampled to reduce the potential bias of abundant values in the low-medium chlorophyll-*a* range, the models did not improve in accurately predicting chlorophyll-*a*. The previously published models also displayed this bias, with no model accurately predicting values of chlorophyll-*a* greater than ~30 µg/L, and following a similar shape as displayed in Fig. 2A. With the lack of correspondence between high observed and predicted chlorophyll-*a* values it seems unlikely a universal Landsat algorithm will accurately predict chlorophyll-*a* concentrations in Oklahoma.

The phycocyanin satellite models performed better than the chlorophyll-*a* models and specifically, the phycocyanin model from Vincent et al. (2004) outperformed our model (Table.

3). The better performance by the phycocyanin models was a surprising finding considering current Landsat sensor bands are not optimized for detecting phycocyanin. This pigment, found only in cyanobacteria, is characterized by an absorption trough around 621 nm (Almuhtaram et al., 2021). Due to the lack of a specific band for phycocyanin, cyanobacteria are detected using bands associated with chlorophyll-*a* when using Landsat satellites, making cyanobacteria frequently indistinguishable from aquatic macrophytes (Oyama et al., 2015). The detection of cyanobacteria is also confounded by turbidity, as bands commonly used in chlorophyll-*a* retrieval algorithms are susceptible to interference by suspended sediment (Almuhtaram et al., 2021; Shi et al., 2019). In fact, others have hypothesized the Vincent et al. (2004) algorithm is simply detecting turbidity or chlorophyll-*a* which are correlated with phycocyanin (Hunter et al., 2010). The phycocyanin dataset (N = 97) was an order of magnitude smaller than chlorophyll-*a* dataset and only represented two lakes, Lake Thunderbird and Lake Texoma. The small number of lakes could contribute to the increased predictability in these models, potentially due to a decrease in noise by fewer optical characteristic combinations in the data, as discussed above. Since our phycocyanin model outperformed our chlorophyll-*a* model, we suggest implementing the phycocyanin model if satellites are the only possible source of spectral data. Unlike chlorophyll-*a*, phycocyanin has been shown to be well correlated with cyanobacterial biomass (Thomson-Lang 2020).

As for our random forest models built using satellite data, the chlorophyll-*a* random forest regression model performed better than all of the literature models but slightly worse than our satellite linear model. The phycocyanin random forest model was better than our model but not better than the best phycocyanin model, the Vincent model. We thought because random forests are based on non-linear models that they would outperform the multiple linear regression

models in our dataset, but this was not the case. While random forest modeling is a newer, machine learning approach that might be touted as a better solution to model problems, one of the drawbacks is the difficulty of implementation of the model, as you cannot 'see' the model or report a formula describing the forest. Trying more 'sophisticated' modeling techniques may not be the correct solution if your underlying data is not showing any predictive trends using simpler modeling methods.

Many of the current standard algorithms were calibrated or validated on very few ground-truthed data points from a small number of lakes. Literature algorithms were calibrated on five or fewer lakes. The number of unique optical characteristics, such as different colors and levels of turbidity and suspended solids, would likely increase with lake number, thus increasing the overall variability in satellite reflectance values. One of the obvious goals of universal algorithms is to predict cyanoHABs across many different lakes and seasons, therefore we think it is important to include a wide range of training data for model building. The lack of representative training data could be why the published satellite-based algorithms, for which training data does not include data from Oklahoma lakes, perform worse than our model on Oklahoma lakes. Given the failure of satellite-based models in predicting chlorophyll-*a* within a single U.S. state, we are doubtful that a truly universal model can be developed.

The disconnect between satellite observations and actual pigment concentrations could also be due to satellite sensors not capturing all water-leaving reflectance due to atmospheric interference or cloud-cover, and not capturing data below the surface of the water column, because the satellite sensors only measure chlorophyll-*a* at the surface, but *in-situ* measurements of chlorophyll-*a* are generally taken across the photic zone for mid-column bloom formers. This would cause mid-column blooming cyanobacteria or well mixed blooms to appear less

concentrated or less pigmented than surface blooms (Coffer et al., 2021a). In Oklahoma lakes, for example, we frequently experience *Raphidiopsis* blooms and other non-surface bloomers (Antunes et al., 2015). These problems would be less of an issue for ground-based sensors that are not affected by atmospheric aerosols and when suspended close to the water surface can capture a large majority of the water leaving reflectance.

Our ground-based multispectral models reliably predicted both chlorophyll-*a* and phycocyanin in Oklahoma lakes and performed remarkably better than the satellite models. The chlorophyll-*a* six-band ratio model we developed slightly outperformed Randolph et al.'s (2008) nested-band ratio chlorophyll-*a* model, with both models having similar R^2 values. The CI and CI-multi performed substantially worse than our model and Randolph et al.'s (2008) model. All four models generally underpredicted at very high levels of chlorophyll-*a*. Our phycocyanin seven-band ratio model outperformed the phycocyanin nested-band ratio model from Randolph et al. (2008), the CI model, and CI-multi model. Notably, our model performed better on extreme phycocyanin values (> 50 µg/L) compared with the established model which had systematic underprediction of high phycocyanin events. While Coffer et al. (2021b, 2021a, 2020) and Handler et al. (2023) have successfully used MERIS and OLCI data to detect cyanobacteria in larger Oklahoma reservoirs, we did not corroborate these results with our simulated data. The CI-multi includes a filtering step with the aim of removing non-cyanobacterial blooms before calculating the CI (Coffer et al. 2020). This filtering step seemed to fail in our systems, as many samples with low to moderate phycocyanin concentrations were incorrectly scored as 0 µg/L. The ability to predict low phycocyanin concentrations is valuable because it allows us to detect blooms as they develop. If detected early, measures may be taken to lower risks to humans, like closing swim beaches and posting educational signs. Reporting false negatives (model returning

no bloom when blooms are present) could lead to failure to act to protect the public, pets, and livestock.

In Oklahoma, we found that ground-based instruments show more promise than satellite-based instruments for monitoring cyanoHABs, but other methods exist for cyanobacteria bloom detection with different accuracy, scale, and cost tradeoffs. These alternative approaches include various satellite sensors, ground-based remote sensing, airplane-based sensors, and fluorometry options. Satellites with freely available images excel in the category of low cost, but as we have demonstrated, they lack accuracy and do not sample at spatial scales relevant to small or dendritic reservoirs (Fig. S6). Additionally, clouds are often an underestimated problem in collecting regular satellite images, which may limit timely detection of algal blooms (Ju and Roy 2008). As we have shown, ground-based remote sensing offers high accuracy at a fine scale but has an upfront cost for multi- or hyper-spectral sensors. As a bonus, such sensors could be deployed aerially, giving broader coverage. For example, the National Ecological Observatory Network (NEON) has deployed airplane-based hyperspectral sensors at NEON sites to measure land-cover and vegetation metrics at a fine scale without the interference of clouds (https://www.neonscience.org/data-collection/imaging-spectrometer). Deployed sondes have high up-front and maintenance costs but provide highly accurate measurements on a fine scale. Manual sampling of water followed by in-lab extraction of chlorophyll-*a* has a high cost of human-hours but is accurate and operates on a fine scale (Almuhtaram et al. 2021). Given our results, we suggest that ground-based remote sensing offers the best of both worlds in terms of cost, accuracy, and scale for many scenarios, however the needs of the researcher or water quality manager will determine which of these approaches best meet their needs based on cost, accuracy, and scale.

Our finding that models using ground-based multispectral data better predict chlorophyll-*a* and phycocyanin concentrations compared to models built using Landsat imagery is likely applicable to small lakes and reservoirs around the globe, though transferring models such as ours to other systems may require additional ground-truth data and model adjustment. That is, we do not propose our model as a 'universal model', but rather we suggest our approach of ground-based remote sensing coupled with custom spectral bands and band ratios. The lakes we analyzed in Oklahoma range from 49 to 42,695 hectares in size (Table S1) and are representative of smaller lakes and reservoirs across the southern Great Plains. There are millions of small lakes in the world (Verpoorter et al. 2014, Cael & Seekell 2016) and reservoirs cover approximately 0.26 Mkm^2 globally (Downing et. al 2006). Many of these small waterbodies experience cyanoHABs that cannot be clearly resolved by larger satellite pixels. Because of this resolution limitation, small lakes and reservoirs are likely underrepresented in remote sensing monitoring programs. While Landsat has the appropriate pixel size for sensing small waterbodies, it does not have a band useful for detecting phycocyanin. As previously concluded by Beck et al. (2016), future satellite-based remote sensing for cyanoHABs will require higher resolution (30-m pixels or less), and more appropriate bands that are narrower and similar to those of WorldView-2/-3 and Sentinel-3. In the meantime, ground-based sensors offer an excellent alternative for small- to mid-sized lakes.

5. CONCLUSION

Based on a multi-year, multi-lake comparison of *in-situ* algal pigment data with Landsat- and ground-derived reflectance models, we conclude that:

- Oklahoma-derived Landsat- and ground-based models outperform established reflectance-pigment models for Oklahoma reservoirs.
- Ground-based, multispectral models are superior to Landsat-based models for predicting cyanoHABs in Oklahoma reservoirs.
- Ground-based, multispectral solutions can offer cost-efficient cyanoHAB monitoring in small- to mid-sized systems where satellites may not be appropriate.

ACKNOWLEDGMENTS

Landsat Surface Reflectance products are courtesy of the U.S. Geological Survey Earth Resources Observation and Science Center. Special thanks to: Karen Glenn, James Easton, Anne Easton and Thayer Hallidayschult for sample collection and processing; Brandon Chien, Dani Glidewell, Lena Wilson, Mary Larson, and Garrett Eakers for image collection and processing; the University of Oklahoma Biological Station staff for facilities support; Jie Wang and Jinheng Zhang for sample collection and data processing, and Rajen Bajgain, Yuting Zhou, and Zhenhua Zou for technical assistance; Julie Chambers and Chris Adams for providing pigment data; Rich Zamor and Steve Nikolai for sampling assistance and providing data; Andy Dzialowski and Keith Loftin for early discussions. This work will be submitted by KVC as partial fulfillment of the requirements for a Ph.D. degree from the University of Oklahoma; additional thanks to her committee members, Drs. Alan Wilson, Katharine Marske, and Amy McGovern, for their support in the preparation of this manuscript. This research was supported in part by grants from the Oklahoma Water Resources Research Institute and the Oklahoma Water Resources Board (Agreement No. 2013OK292B-OU1/2), the Oklahoma Department of Wildlife

Conservation (through the Sport Fish Restoration Program, Grant F-61-R), NSF (DEB 1831061), USGS (G21AP10181), and an NSF Graduate Research Fellowship.

WORKS CITED

Alawadi, F., 2010. Detection of surface algal blooms using the newly developed algorithm surface algal bloom index (SABI). International Society for Optics and Photonics, p. 782506. https://doi.org/10.1117/12.862096

Almuhtaram, H., Kibuye, F.A., Ajjampur, S., Glover, C.M., Hofmann, R., Gaget, V., Owen, C., Wert, E.C., Zamyadi, A., 2021a. State of knowledge on early warning tools for cyanobacteria detection. Ecological Indicators 133, 108442. https://doi.org/10.1016/j.ecolind.2021.108442

Almuhtaram, H., Kibuye, F.A., Ajjampur, S., Glover, C.M., Hofmann, R., Gaget, V., Owen, C., Wert, E.C., Zamyadi, A., 2021b. State of knowledge on early warning tools for cyanobacteria detection. Ecological Indicators 133, 108442. https://doi.org/10.1016/j.ecolind.2021.108442

Baban, S.M.J., 1993. Detecting water quality parameters in the Norfolk Broads, U.K., using Landsat imagery. International Journal of Remote Sensing 14, 1247–1267. https://doi.org/10.1080/01431169308953955

Barnes, W.L., Pagano, T.S., Salomonson, V.V., 1998. Prelaunch characteristics of the Moderate Resolution Imaging Spectroradiometer (MODIS) on EOS-AM1. IEEE Transactions on Geoscience and Remote Sensing 36, 1088–1100. https://doi.org/10.1109/36.700993

Beck, R., Zhan, S., Liu, H., Tong, S., Yang, B., Xu, M., Ye, Z., Huang, Y., Shu, S., Wu, Q., Wang, S., Berling, K., Murray, A., Emery, E., Reif, M., Harwood, J., Young, J., Nietch,

C., Macke, D., Martin, M., Stillings, G., Stump, R., Su, H., 2016a. Comparison of satellite reflectance algorithms for estimating chlorophyll-a in a temperate reservoir using coincident hyperspectral aircraft imagery and dense coincident surface observations. Remote Sensing of Environment 178, 15–30. https://doi.org/10.1016/j.rse.2016.03.002

Beck, R., Zhan, S., Liu, H., Tong, S., Yang, B., Xu, M., Ye, Z., Huang, Y., Shu, S., Wu, Q., Wang, S., Berling, K., Murray, A., Emery, E., Reif, M., Harwood, J., Young, J., Nietch, C., Macke, D., Martin, M., Stillings, G., Stump, R., Su, H., 2016b. Comparison of satellite reflectance algorithms for estimating chlorophyll-a in a temperate reservoir using coincident hyperspectral aircraft imagery and dense coincident surface observations. Remote Sensing of Environment 178, 15–30. https://doi.org/10.1016/j.rse.2016.03.002

Beyer, J.E., Hambright, K.D., 2017. Maternal effects are no match for stressful conditions: a test of the maternal match hypothesis in a common zooplankter. Functional Ecology 31, 1933–1940. https://doi.org/10.1111/1365-2435.12901

Brivio, P.A., Giardino, C., Zilioli, E., 2001. Determination of chlorophyll concentration changes in Lake Garda using an image-based radiative transfer code for Landsat TM images. International Journal of Remote Sensing 22, 487–502. https://doi.org/10.1080/014311601450059

Coad, P., Cathers, B., Ball, J.E., Kadluczka, R., 2014. Proactive management of estuarine algal blooms using an automated monitoring buoy coupled with an artificial neural network. Environmental Modelling & Software 61, 393–409. https://doi.org/10.1016/j.envsoft.2014.07.011

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., Foreman, K., Porteous, A., Loftin, K.A., Stumpf, R.P., Werdell, P.J., Urquhart, E., Albert, R.J., Darling, J.A., 2021a. Assessing cyanobacterial frequency and abundance at surface waters near drinking water intakes across the United States. Water Research 201, 117377. https://doi.org/10.1016/j.watres.2021.117377

Coffer, M.M., Schaeffer, B.A., Salls, W.B., Urquhart, E., Loftin, K.A., Stumpf, R.P., Werdell, P.J., Darling, J.A., 2021b. 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

Dall'Olmo, G., Gitelson, A.A., 2005. Effect of bio-optical parameter variability on the remote estimation of chlorophyll-a concentration in turbid productive waters: experimental results. Appl. Opt., AO 44, 412–422. https://doi.org/10.1364/AO.44.000412

Fernandez-Figueroa, E.G., Wilson, A.E., Rogers, S.R., 2021. Commercially available unoccupied aerial systems for monitoring harmful algal blooms: A comparative study. Limnology and Oceanography: Methods. https://doi.org/10.1002/lom3.10477

Gitelson, A.A., Gritz †, Y., Merzlyak, M.N., 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. Journal of Plant Physiology 160, 271–282. https://doi.org/10.1078/0176-1617-00887

Gons, H.J., Rijkeboer, M., Ruddick, K.G., 2002. A chlorophyll-retrieval algorithm for satellite imagery (Medium Resolution Imaging Spectrometer) of inland and coastal waters. J. Plankton Res. 24, 947–951. https://doi.org/10.1093/plankt/24.9.947

Hambright, K.D., Zamor, R.M., Easton, J.D., Allison, B., 2014. Algae, in: Encyclopedia of Toxicology. Elsevier, pp. 130–141. https://doi.org/10.1016/B978-0-12-386454-3.00983-0

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

Hilborn, E.D., Beasley, V.R., 2015. One Health and Cyanobacteria in Freshwater Systems: Animal Illnesses and Deaths Are Sentinel Events for Human Health Risks. Toxins 7, 1374–1395. https://doi.org/10.3390/toxins7041374

Huisman, J., Codd, G.A., Paerl, H.W., Ibelings, B.W., Verspagen, J.M.H., Visser, P.M., 2018. Cyanobacterial blooms. Nat Rev Microbiol 16, 471–483. https://doi.org/10.1038/s41579-018-0040-1

Hunter, P.D., Tyler, A.N., Carvalho, L., Codd, G.A., Maberly, S.C., 2010. Hyperspectral remote sensing of cyanobacterial pigments as indicators for cell populations and toxins in eutrophic lakes. Remote Sensing of Environment 114, 2705–2718. https://doi.org/10.1016/j.rse.2010.06.006

Ju, J., Roy, D.P., 2008. The availability of cloud-free Landsat ETM+ data over the conterminous United States and globally. Remote Sensing of Environment 112, 1196–1211. https://doi.org/10.1016/j.rse.2007.08.011

Kneubühler, M., Frank, T., Kellenberger, T., Pasche, N., Schmid, M., 2007. Mapping chlorophyll-a in Lake Kivu with remote sensing methods, in: Lacoste, H., Ouwehand, L. (Eds.), . European Space Agency * Communication Production Office, Montreux (CH).

Landsat surface reflectance data (Report No. 2015–3034), 2015. , Fact Sheet. Reston, VA. https://doi.org/10.3133/fs20153034

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18–22.

Markham, B.L., Storey, J.C., Williams, D.L., Irons, J.R., 2004. Landsat sensor performance: history and current status. IEEE Transactions on Geoscience and Remote Sensing 42, 2691–2694. https://doi.org/10.1109/TGRS.2004.840720

Masek, J.G., Vermote, E.F., Saleous, N.E., Wolfe, R., Hall, F.G., Huemmrich, K.F., Gao, F., Kutler, J., Lim, T.-K., 2006. A Landsat surface reflectance dataset for North America, 1990-2000. IEEE Geoscience and Remote Sensing Letters 3, 68–72. https://doi.org/10.1109/LGRS.2005.857030

Matthews, M.W., 2011. A current review of empirical procedures of remote sensing in inland and near-coastal transitional waters. International Journal of Remote Sensing 32, 6855–6899. https://doi.org/10.1080/01431161.2010.512947

Matthews, M.W., Bernard, S., Robertson, L., 2012a. An algorithm for detecting trophic status (chlorophyll-a), cyanobacterial-dominance, surface scums and floating vegetation in inland and coastal waters. Remote Sensing of Environment 124, 637–652. https://doi.org/10.1016/j.rse.2012.05.032

Matthews, M.W., Bernard, S., Robertson, L., 2012b. An algorithm for detecting trophic status (chlorophyll-a), cyanobacterial-dominance, surface scums and floating vegetation in inland and coastal waters. Remote Sensing of Environment 124, 637–652. https://doi.org/10.1016/j.rse.2012.05.032

Mishra, S., Mishra, D.R., 2012. Normalized difference chlorophyll index: A novel model for remote estimation of chlorophyll-a concentration in turbid productive waters. Remote

Sensing of Environment, Remote Sensing of Urban Environments 117, 394–406. https://doi.org/10.1016/j.rse.2011.10.016

Nieke, J., Borde, F., Mavrocordatos, C., Berruti, B., Delclaud, Y., Riti, J.B., Garnier, T., 2012. The Ocean and Land Colour Imager (OLCI) for the Sentinel 3 GMES Mission: status and first test results. International Society for Optics and Photonics, p. 85280C.

Oklahoma Water Resources Board, 2017. Beneficial Use Monitoring Program (Lake Report). Oklahoma Water Resources Board, State of Oklahoma, Oklahoma City, OK.

Overstreet, B.T., Legleiter, C.J., 2017. Removing sun glint from optical remote sensing images of shallow rivers. Earth Surface Processes and Landforms 42, 318–333. https://doi.org/10.1002/esp.4063

Oyama, Y., Matsushita, B., Fukushima, T., 2015. Distinguishing surface cyanobacterial blooms and aquatic macrophytes using Landsat/TM and ETM+ shortwave infrared bands. Remote Sensing of Environment, Special Issue: Remote Sensing of Inland Waters 157, 35–47. https://doi.org/10.1016/j.rse.2014.04.031

Paerl, H.W., Barnard, M.A., 2020. Mitigating the global expansion of harmful cyanobacterial blooms: Moving targets in a human- and climatically-altered world. Harmful Algae 96, 101845. https://doi.org/10.1016/j.hal.2020.101845

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

Plaas, H.E., Paerl, H.W., 2021. Toxic Cyanobacteria: A Growing Threat to Water and Air Quality. Environ. Sci. Technol. 55, 44–64. https://doi.org/10.1021/acs.est.0c06653

R Development Core Team, 2014. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Randolph, K., Wilson, J., Tedesco, L., Li, L., Pascual, D.L., Soyeux, E., 2008. Hyperspectral remote sensing of cyanobacteria in turbid productive water using optically active pigments, chlorophyll a and phycocyanin. Remote Sensing of Environment, Applications of Remote Sensing to Monitoring Freshwater and Estuarine Systems 112, 4009–4019. https://doi.org/10.1016/j.rse.2008.06.002

Rast, M., Bezy, J.L., Bruzzi, S., 1999. The ESA Medium Resolution Imaging Spectrometer MERIS a review of the instrument and its mission. International Journal of Remote Sensing 20, 1681–1702. https://doi.org/10.1080/014311699212416

Schwarz, G., 1978. Estimating the Dimension of a Model. Ann. Statist. 6, 461–464. https://doi.org/10.1214/aos/1176344136

Shi, K., Zhang, Y., Qin, B., Zhou, B., 2019. Remote sensing of cyanobacterial blooms in inland waters: present knowledge and future challenges. Science Bulletin 64, 1540–1556. https://doi.org/10.1016/j.scib.2019.07.002

Stroming, S., Robertson, M., Mabee, B., Kuwayama, Y., Schaeffer, B., 2020. Quantifying the Human Health Benefits of Using Satellite Information to Detect Cyanobacterial Harmful Algal Blooms and Manage Recreational Advisories in U.S. Lakes. GeoHealth 4, e2020GH000254. https://doi.org/10.1029/2020GH000254

Thomson-Laing, G., Puddick, J., Wood, S.A., 2020. Predicting cyanobacterial biovolumes from phycocyanin fluorescence using a handheld fluorometer in the field. Harmful Algae 97, 101869. https://doi.org/10.1016/j.hal.2020.101869

Tomlinson, M.C., Stumpf, R.P., Wynne, T.T., Dupuy, D., Burks, R., Hendrickson, J., III, R.S.F., 2016. Relating chlorophyll from cyanobacteria-dominated inland waters to a MERIS

bloom index. Remote Sensing Letters 7, 141–149. https://doi.org/10.1080/2150704X.2015.1117155

Urquhart, E.A., Schaeffer, B.A., Stumpf, R.P., Loftin, K.A., Werdell, P.J., 2017. A method for examining temporal changes in cyanobacterial harmful algal bloom spatial extent using satellite remote sensing. Harmful Algae 67, 144–152. https://doi.org/10.1016/j.hal.2017.06.001

Verpoorter, C., Kutser, T., Seekell, D.A., Tranvik, L.J., 2014. A global inventory of lakes based on high-resolution satellite imagery. Geophysical Research Letters 41, 6396–6402. https://doi.org/10.1002/2014GL060641

Vincent, R.K., Qin, X., McKay, R.M.L., Miner, J., Czajkowski, K., Savino, J., Bridgeman, T., 2004. Phycocyanin detection from LANDSAT TM data for mapping cyanobacterial blooms in Lake Erie. Remote Sensing of Environment 89, 381–392. https://doi.org/10.1016/j.rse.2003.10.014

Wetzel, R.G., 2001. Limnology: lake and river ecosystems. Gulf Professional Publishing.

Wu, D., Li, R., Zhang, F., Liu, J., 2019. A review on drone-based harmful algae blooms monitoring. Environ Monit Assess 191, 211. https://doi.org/10.1007/s10661-019-7365-8

Wynne, T.T., Stumpf, R.P., Tomlinson, M.C., Dyble, J., 2010. Characterizing a cyanobacterial bloom in Western Lake Erie using satellite imagery and meteorological data. Limnology and Oceanography 55, 2025–2036. https://doi.org/10.4319/lo.2010.55.5.2025

Zhao, D., Xing, X., Liu, Y., Yang, J., Wang, L., 2010. The relation of chlorophyll-a concentration with the reflectance peak near 700 nm in algae-dominated waters and sensitivity of fluorescence algorithms for detecting algal bloom. International Journal of Remote Sensing 31, 39–48. https://doi.org/10.1080/01431160902882512

Zou, Z., Dong, J., Menarguez, M.A., Xiao, X., Qin, Y., Doughty, R.B., Hooker, K.V., David Hambright, K., 2017. Continued decrease of open surface water body area in Oklahoma during 1984–2015. Science of The Total Environment 595, 451–460. https://doi.org/10.1016/j.scitotenv.2017.03.259

Zuur, A.F., Ieno, E.N., Elphick, C.S., 2010. A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution 1, 3–14. https://doi.org/10.1111/j.2041-210X.2009.0000

TABLES

Table 1. Band math for each algorithm (derived from Beck et al. 2016, Vincent et al. 2004, and Randolph et al. 2008) used to estimate chlorophyll-*a* (1-5, 7, 9-10) and phycocyanin (6, 8-10) concentrations in Oklahoma lakes and each original citation. In addition, we have listed the type of algorithm and the original location and the algorithm for the relevant sensor. Ground-truthed data refers to pigment concentrations from waterbodies that were used to calibrate (build the model) or validate (test the model). Please see equations 4 and 5 in Randolph et al. (2008) for complete algorithms.

	Algorithm	Equation	Citation	Type, location	Original sensor	# Ground-truthed data	Ground-truthed waterbody	Method
1	NDCI	(b4 - b3)/(b4 + b3)	Mishra & Mishra 2012	semi-empirical, oceans	MERIS	56	Chesapeake Bay, Delaware Bay, the river Mississippi Delta region, and the Mobile Bay, USA	Calibrated with simulated data, validated with field data
2	SABI	(b4 - b3)/(b1 + b2)	Alawadi 2010	semi-empirical, surface blooms	MODIS	0	--	Used Chlor-a MODIS product to validate
3	FLH blue	(b2)-[b3 + (b1 - b3)]	Zhao et al. 2010	semi-empirical, oceans	Hyperspectral spectroradio-meter	41	--	Validated with laboratory cultures
4	3BDA	(b3 - b4)*b4	Dall'Olmo & Gitelson 2005	semi-empirical, reservoirs	Hyperspectral radiometer	144	2 sand pit lakes and 2 reservoirs, Nebraska; 1 lake, Iowa, USA	Calibrated with N = 86 and validation N = 58
5	KIVU (3BDA – like)	(b1 - b3)/(b2)	Brivio et al. 2001	semi-empirical, lakes	Landsat 5 TM	6	Lake Garda, Italy	Validated with in-situ data
6	Vincent	(b3/ b1) + (b4/b1) + (b4/b3) + (b5/b3) + (b7/b3) + (b7/b4)	Vincent et al. 2004	empirical, great lakes	Landsat 7 ETM+ and Landsat 5 TM	52	Lake Erie, USA	Calibrated model with in-situ data
7	Randolph (chla)	$(((R(709)/R(620))*(0.727 + b_b)) – b_b – 0.401) * (1/0.68)$	Randolph et al. 2008	semi-empirical, reservoirs	Hyperspectral spectroradio-meter	55	Geist and Morse reservoirs, Indiana, USA	Calibrated model with in-situ data
8	Randolph (pcy)	$((((R(709)/R(620))*(0.727 + b_b)) – b_b – 0.281) * (1/0.84)) – (0.24*a_{chl})$	Randolph et al. 2008	semi-empirical, reservoirs	Hyperspectral spectroradio-meter	55	Geist and Morse reservoirs, Indiana, USA	Calibrated model with in-situ data
9	Cyanobacteria Index (CI)	$SS(\lambda) = R(\lambda) – R(\lambda') – [R(\lambda'') – R(\lambda')] * [(\lambda - \lambda')/(\lambda'' – \lambda')]$ $(CI) = -SS(\lambda)$	Wynne et al. (2008)	empirical, lakes	Hyperspectral spectroradio-meter, MERIS	NA	Bear Lake, MI, USA; Saginaw Bay, Lake Huron, MI, USA	Calibrated model with in-situ spectra, validated with satellite imagery
10	Cyanobacteria Index-multi (CI-multi)	If $-SS(665) < 0$, CI-multi = 0 If $-SS(665) > 0$, CI-multi = CI	Matthews et al. (2012)	empirical, oceans & lakes/reservoirs	MERIS	74	Benguela (Atlantic Ocean), Loskop Dam Reservoir (South Africa), Zeekoevlei (South Africa), Hartbeespoort Dam Reservoir (South Africa)	Validated with in-situ data

Table 2. List of bands used to build a semi-empirical model to estimate phycocyanin. Bands were developed from hyperspectral data collected from Oklahoma lakes.

Band	Purpose	Minimum reflectance (nm)	Maximum reflectance (nm)
1	chlorophyll absorption minimum	546	591
2	phycocyanin trough	610	620
3	chlorophyll absorption trough	660	670
4	chlorophyll trough	678	684
5	reference band	700	710

Table 3. Performance of satellite algorithms for chlorophyll-*a* and phycocyanin estimation on Oklahoma lakes was evaluated using the adjusted R^2 from the given linear model, the root mean square error (RMSE), and Pearson's r correlation coefficient. *The RMSE for the resampled model should not be directly compared to the other models because the dataset used to build the model was different.

	Algorithm name	Adjusted R^2	Model P-value	RMSE	Pearson's r
	Chlorophyll-*a* algorithms				
1	NDCI	-0.0009	0.81	17.57	0.007
2	SABI	0.001	0.16	17.55	0.044
3	FLHB	-0.0009	0.90	17.57	0.004
4	2BDA	0.002	0.10	17.55	0.050
5	3BDA	0.0003	0.25	17.56	0.035
6	KIVU (3BDA – like)	0.003	0.05	17.54	0.060
7	this paper (full dataset)	0.035	<0.001	17.22	0.198
8	this paper (resampled dataset)	0.025	<0.001	21.95*	0.170
	Phycocyanin algorithms				
9	Vincent	0.384	<0.001	16.85	0.650
10	this paper	0.291	<0.001	18.49	0.553
	Random Forests	**R2**	**% Variance explained**	**RMSE**	**Model significantly different from null?**
11	chlorophyll-*a*	0.011	3.07	17.30	No
12	phycocyanin	0.321	32.86	18.18	No

Table 4. Performance of hyperspectral algorithms for chlorophyll-*a* and phycocyanin estimation on Oklahoma lakes evaluated using the adjusted R^2 from the given linear model, the root mean square error (RMSE), and Pearson's r correlation coefficient.

	Algorithm name	Adjusted R^2	Model P-value	RMSE	Pearson's r
	Chlorophyll-*a* algorithms				
1	Randolph et al. model	0.509	<0.001	22.07	0.717
2	CI	0.472	<0.001	20.26	0.690
3	CI-multi	0.376	<0.001	21.93	0.617
3	this paper	0.660	<0.001	16.03	0.814
	Phycocyanin algorithms				
5	Randolph et al. model	0.592	<0.001	34.99	0.772
6	CI	0.594	<0.001	69.33	0.773
7	CI-multi	0.519	<0.001	69.33	0.723
8	this paper	0.816	<0.001	19.16	0.904

FIGURES

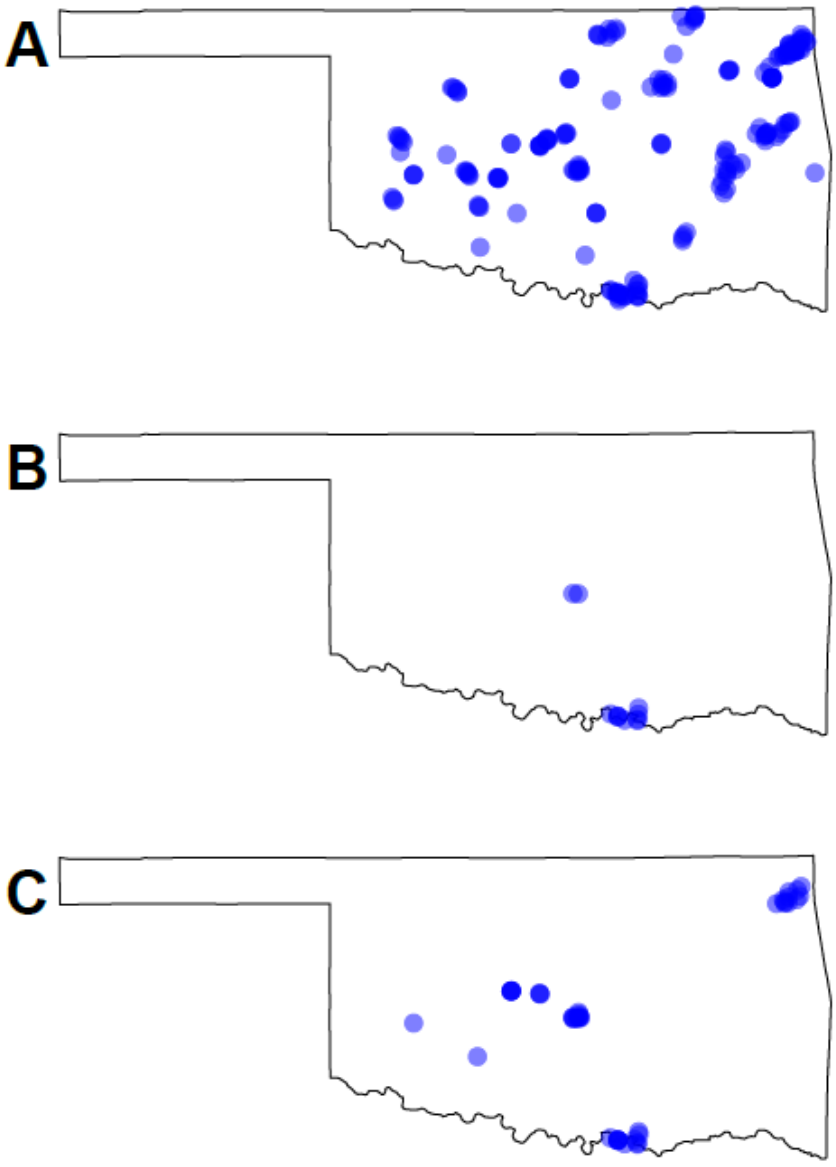


Figure 1. Map of study sites in the US state of Oklahoma for (A) satellite-based chlorophyll-a model (N = 191), (B) satellite-based phycocyanin model (N = 11), and (C) ground-based multispectral chlorophyll-a and phycocyanin models (N = 45). Each point represents one site. Darker colors indicate overlap of geographically close sites.

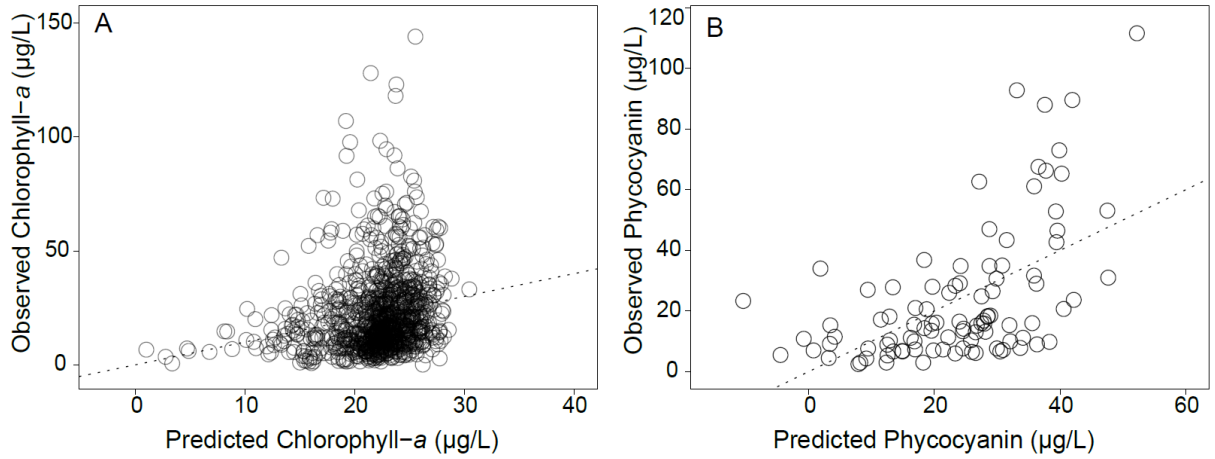


Figure 2. Comparison of predicted and observed in-situ concentrations of algal pigments based on models built using Landsat 7 and 8 reflectance values. (A) The best chlorophyll-a (N = 1060) model based on BIC had an adjusted R2 = 0.035 and $p \leq 0.001$. (B) The best phycocyanin (N = 97) model based on BIC had an adjusted R2 = 0.29 and $p \leq 0.001$. Dashed line shows a 1:1 ratio between observed and predicted pigment concentrations. Darker areas in (A) indicate increased overlap of points.

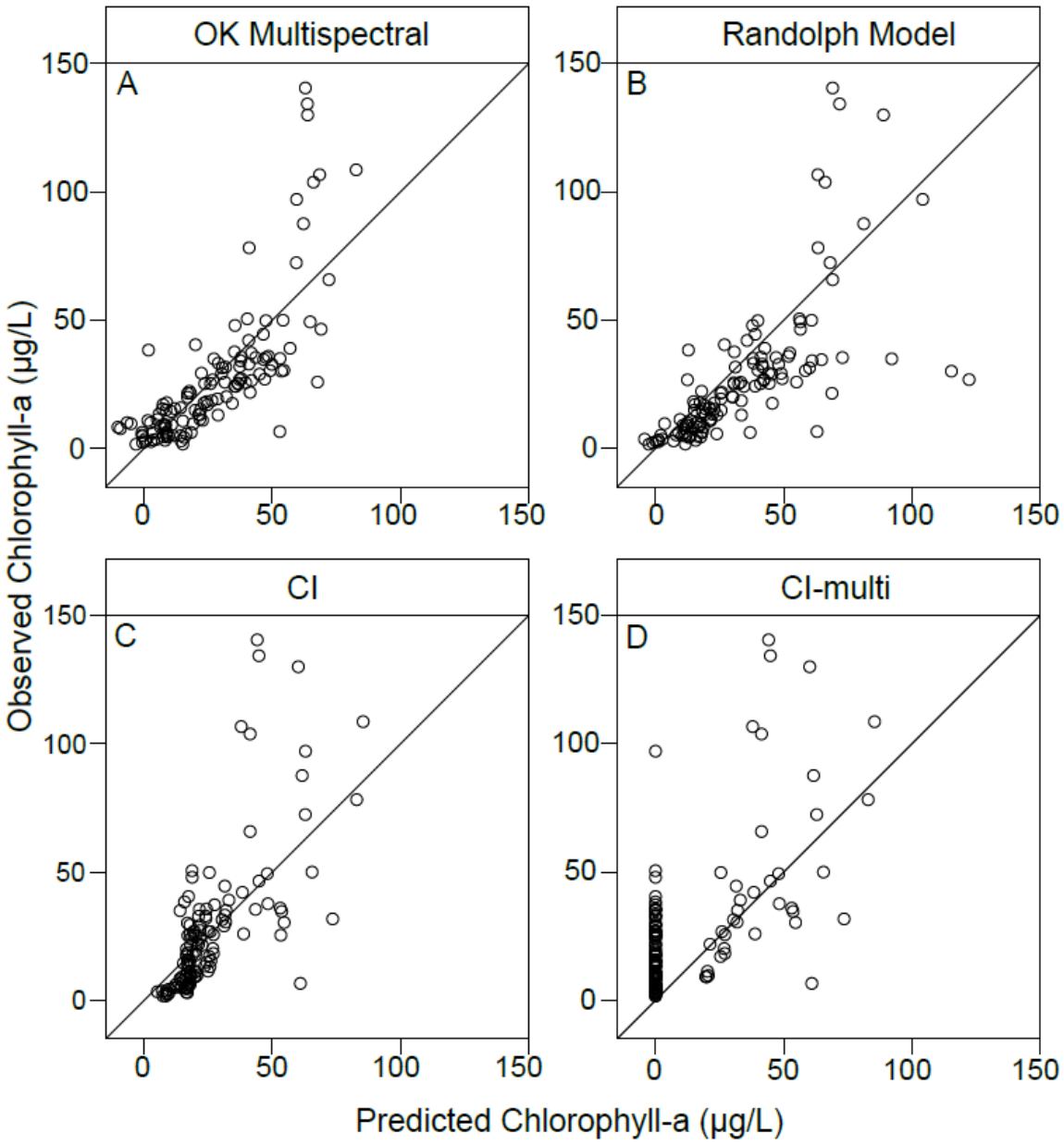


Figure 3. Comparison of the observed in-situ chlorophyll-a and the predicted chlorophyll-a concentrations from (A) the multispectral 6-band ratio model from this paper, (B) the Randolph model, (C) the Cyanobacterial Index (CI), and (D) the updated Cyanobacterial Index (CI-multi). See Table 4 for model fit statistics. The solid line shows a 1:1 ratio between observed and predicted pigment concentrations to visually show the accuracy of the predicted values (closer to the line is more accurate). Darker areas indicate increased overlap of points.

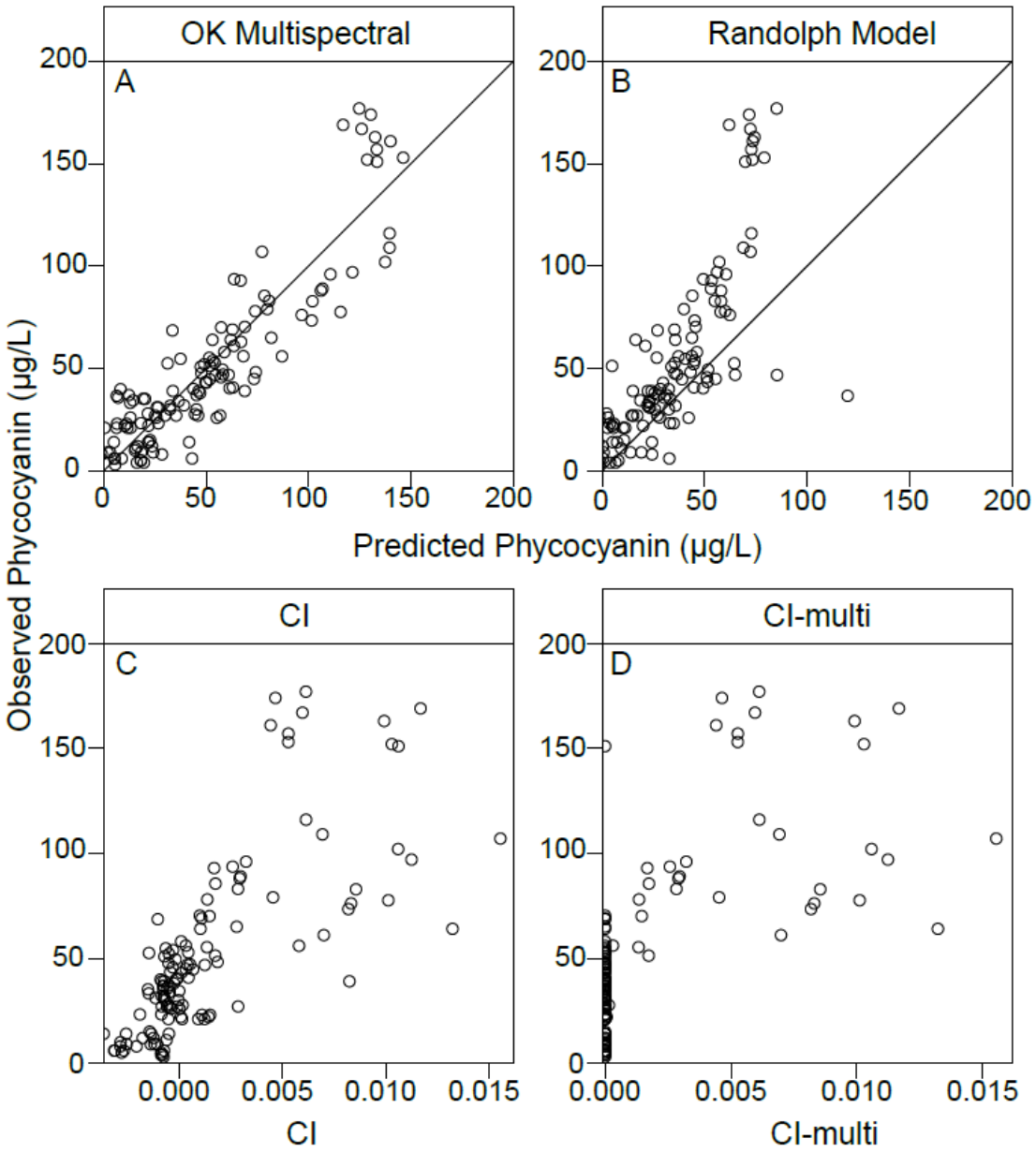


Figure 4. Comparison of the observed and predicted concentrations of phycocyanin based on the multispectral 7-band ratio model from this paper (A), the Randolph model (B), CI, and CI-multi. Model fit statistics can be found in Table 4. The CI (C) and CI-multi (D) were used to 'predict' phycocyanin. There are currently no models translating the raw model output (CI, CI-multi) to µg/L phycocyanin, therefore observed phycocyanin was plotted against the raw model outputs (unitless) to test correlation. The solid line in panels A and B shows a 1:1 ratio between observed

and predicted pigment concentrations to visually show the accuracy of the predicted values (closer to the line is more accurate). Darker areas indicate increased overlap of points.

Appendix 1

Supplemental Materials to Chapter 1:

The global *Microcystis* interactome

Katherine V. Cook, C. Li, H. Cai, L.R. Krumholz, K.D. Hambright, H.W. Paerl, M.M. Steffen, A.E. Wilson, M. Burford, H.-P. Grossart, D. P. Hamilton, H. Jiang, A. Sukenik, D. Latour, E.I. Meyer, J. Padisák, B. Qin, R.M. Zamor, and G. Zhu

---

Supplementary Table 1. Summary of metagenome data for each lake.

Lake	category	contigs number	contigs bases(Mbp)	N50(bp)	Max length(bp)	protein-coding gene number
Belső-tó	Bacteria	33496	27.1	765	74324	26699
	Microcystis	890	4.7	16388	81696	4267
Chaohu	Bacteria	23682	19.8	794	36145	18385
	Microcystis	5258	7.8	1944	54606	6268
Clarendon	Bacteria	82636	104.8	1392	321729	102844
	Microcystis	6740	10.9	2336	38246	9427
FP23	Bacteria	155545	19.5	1454	49139	28976
	Microcystis	4816	7.9	2248	35883	6825
Grand	Bacteria	73551	92.3	1418	105344	89141
	Microcystis	9312	11.1	1418	57431	8881
Kinneret	Bacteria	26202	37.0	1817	101729	37307
	Microcystis	4943	9.1	2627	55640	8028
Aasee	Bacteria	36892	57.9	2441	299862	56457
	Microcystis	1122	5.5	19101	77987	5191
Rotoehu	Bacteria	26009	31.1	1368	41132	32313
	Microcystis	503	4.3	16710	49814	4098
Villerest	Bacteria	30454	40.5	1691	97662	36391
	Microcystis	2619	6.7	5702	54900	5927

Supplementary Table 2. KEGG orthology numbers and functional pathways found in *Microcystis* and the microbiome bacteria, indicating involvement in carbon (C), nitrogen (N), phosphorus (P), or sulfur (S) cycling. Complete pathways are indicated by black fill while modules with no more than one pathway missing, are indicated by blue fill . Empty cells indicate missing or partial (>1 missing) pathways.

Pathway	*Microcystis*	Bacteria	Biogeochemical cycle
Carbohydrate metabolism			
M00001 Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate			**C**
M00002 Glycolysis, core module involving three-carbon compounds			**C**
M00003 Gluconeogenesis, oxaloacetate => fructose-6P			**C, P**
M00307 Pyruvate oxidation, pyruvate => acetyl-CoA			**C**
M00009 Citrate cycle (TCA cycle, Krebs cycle)			**C**
M00010 Citrate cycle, first carbon oxidation, oxaloacetate => 2-oxoglutarate			**C, S**
M00011 Citrate cycle, second carbon oxidation, 2-oxoglutarate => oxaloacetate			**C**
M00004 Pentose phosphate pathway (Pentose phosphate cycle)			**C, P**
M00006 Pentose phosphate pathway, oxidative phase, glucose 6P => ribulose 5P			**C,P**
M00007 Pentose phosphate pathway, non-oxidative phase, fructose 6P => ribose 5P			**C,P**
M00580 Pentose phosphate pathway, archaea, fructose 6P => ribose 5P			**C,P**
M00005 PRPP biosynthesis, ribose 5P => PRPP			**C,P**
M00008 Entner-Doudoroff pathway, glucose-6P => glyceraldehyde-3P + pyruvate			**C,P**
M00308 Semi-phosphorylative Entner-Doudoroff pathway, gluconate => glycerate-3P			**C,P**

M00309 Non-phosphorylative Entner-Doudoroff pathway, gluconate/galactonate => glycerate	**C**
M00854 Glycogen biosynthesis, glucose-1P => glycogen/starch	**C,P**
M00855 Glycogen degradation, glycogen => glucose-6P	**C,P**
M00565 Trehalose biosynthesis, D-glucose 1P => trehalose	**C,P**
M00549 Nucleotide sugar biosynthesis, glucose => UDP-glucose	**C,P**
M00631 D-Galacturonate degradation, D-galacturonate => pyruvate + D glyceraldehyde 3P	**C,P**
M00061 D-Glucuronate degradation, D-glucuronate => pyruvate + D-glyceraldehyde 3P	**C,P**
M00632 Galactose degradation, Leloir pathway, galactose => alpha-D-glucose-1P	**C,P**
M00552 D-galactonate degradation, De Ley-Doudoroff pathway, D-galactonate => glycerate-3P	**C,P**
M00554 Nucleotide sugar biosynthesis, galactose => UDP-galactose	**C,P**
M00012 Glyoxylate cycle	**C**
M00373 Ethylmalonyl pathway	**C**
M00013 Malonate semialdehyde pathway, propanoyl-CoA => acetyl-CoA	**C**
M00741 Propanoyl-CoA metabolism, propanoyl-CoA => succinyl-CoA	**C**
M00130 Inositol phosphate metabolism, PI=> PIP2 => Ins(1,4,5)P3 => Ins(1,3,4,5)P4	**C,P**
M00132 Inositol phosphate metabolism, Ins(1,3,4)P3 => phytate	**C**

Energy metabolism

Carbon fixation

M00165 Reductive pentose phosphate cycle (Calvin cycle)	**C**
M00166 Reductive pentose phosphate cycle, ribulose-5P => glyceraldehyde-3P	**C,P**
M00167 Reductive pentose phosphate cycle, glyceraldehyde-3P => ribulose-5P	**C,P**

M00168 CAM (Crassulacean acid metabolism), dark	**C**
M00173 Reductive citrate cycle (Arnon-Buchanan cycle)	**C**
M00579 Phosphate acetyltransferase-acetate kinase pathway, acetyl-CoA => acetate	**C,P**
Methane metabolism	
M00345 Formaldehyde assimilation, ribulose monophosphate pathway	**C**
M00358 Coenzyme M biosynthesis	**C**
M00174 Methane oxidation, methanotroph, methane => formaldehyde	**C**
M00346 Formaldehyde assimilation, serine pathway	**C**
M00344 Formaldehyde assimilation, xylulose monophosphate pathway	**C,P**
M00378 F420 biosynthesis	**C**
Nitrogen metabolism	
M00531 Assimilatory nitrate reduction, nitrate => ammonia	**N**
M00175 Nitrogen fixation, nitrogen => ammonia	**N**
M00530 Dissimilatory nitrate reduction, nitrate => ammonia	**N**
M00529 Denitrification, nitrate => nitrogen	**N**
M00528 Nitrification, ammonia => nitrite	**N**
M00804 Complete nitrification, comammox, ammonia => nitrite => nitrate	**N**
Sulfur metabolism	
M00176 Assimilatory sulfate reduction, sulfate => H2S	S
M00595 Thiosulfate oxidation by SOX complex, thiosulfate => sulfate	**S**
M00596 Dissimilatory sulfate reduction, sulfate => H2S	**S**
Photosynthesis	
M00161 Photosystem II	C
M00163 Photosystem I	C
M00597 Anoxygenic photosystem II	**C**
ATP synthesis (Structural complex)	

M00144 NADH:quinone oxidoreductase, prokaryotes

M00146 NADH dehydrogenase (ubiquinone) 1 alpha subcomplex

M00145 NAD(P)H:quinone oxidoreductase, chloroplasts and cyanobacteria

M00149 Succinate dehydrogenase, prokaryotes

M00162 Cytochrome b6f complex

M00154 Cytochrome c oxidase

M00155 Cytochrome c oxidase, prokaryotes

M00153 Cytochrome bd ubiquinol oxidase

M00157 F-type ATPase, prokaryotes and chloroplasts

M00150 Fumarate reductase, prokaryotes

M00162 Cytochrome b6f complex

M00151 Cytochrome bc1 complex respiratory unit

M00152 Cytochrome bc1 complex

M00154 Cytochrome c oxidase

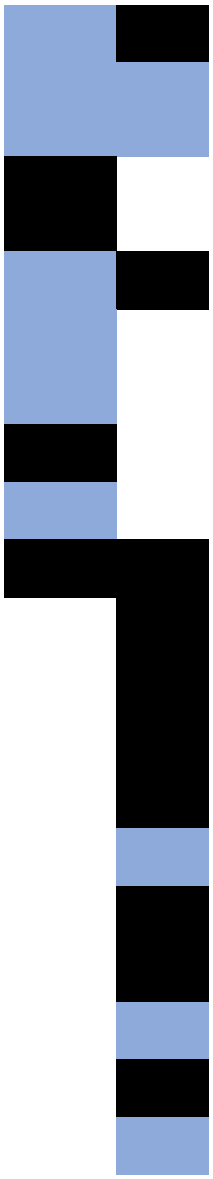
M00155 Cytochrome c oxidase, prokaryotes

M00153 Cytochrome bd ubiquinol oxidase

M00417 Cytochrome o ubiquinol oxidase

M00156 Cytochrome c oxidase, cbb3-type

M00159 V-type ATPase, prokaryotes



Lipid metabolism

Fatty acid metabolism

M00082 Fatty acid biosynthesis, initiation **C**

M00083 Fatty acid biosynthesis, elongation **C**

M00086 beta-Oxidation, acyl-CoA synthesis **C**

M00087 beta-Oxidation **C**

M00861 beta-Oxidation, peroxisome, VLCFA **C**

Sterol biosynthesis

M00862 beta-Oxidation, peroxisome, tri/dihydroxycholestanoyl-CoA => choloyl/chenodeoxycholoyl-CoA	**C**
M00107 Steroid hormone biosynthesis, cholesterol => prognenolone => progesterone	**C**
M00110 C19/C18-Steroid hormone biosynthesis, pregnenolone => androstenedione => estrone	**C**
Lipid metabolism	
M00098 Acylglycerol degradation	C
M00088 Ketone body biosynthesis, acetyl-CoA => acetoacetate/3-hydroxybutyrate/acetone	**C**
M00089 Triacylglycerol biosynthesis	**C**
M00090 Phosphatidylcholine (PC) biosynthesis, choline => PC	**C,P**
M00091 Phosphatidylcholine (PC) biosynthesis, PE => PC	**C,P**
M00092 Phosphatidylethanolamine (PE) biosynthesis, ethanolamine => PE	**C,P**
M00093 Phosphatidylethanolamine (PE) biosynthesis, PA => PS => PE	**C,P**
M00094 Ceramide biosynthesis	**C**
M00066 Lactosylceramide biosynthesis	**C**
M00099 Sphingosine biosynthesis	**C**
M00100 Sphingosine degradation	**C**
Nucleotide metabolism	
Purine metabolism	
M00048 Inosine monophosphate biosynthesis, PRPP + glutamine => IMP	**C,P**
M00049 Adenine ribonucleotide biosynthesis, IMP => ADP,ATP	**C,P**
M00050 Guanine ribonucleotide biosynthesis IMP => GDP,GTP	**C,P**
M00546 Purine degradation, xanthine => urea	**C,N**
Pyrimidine metabolism	

M00051 Uridine monophosphate biosynthesis, glutamine (+ PRPP) => UMP **C,P**

M00052 Pyrimidine ribonucleotide biosynthesis, UMP => UDP/UTP,CDP/CTP **C,P**

M00053 Pyrimidine deoxyribonuleotide biosynthesis, CDP/CTP => dCDP/dCTP,dTDP/dTTP **C,P**

M00046 Pyrimidine degradation, uracil => beta-alanine, thymine => 3-aminoisobutanoate **C,P**

Amino acid metabolism

Serine and threonine metabolism

M00018 Threonine biosynthesis, aspartate => homoserine => threonine **C,N**

M00555 Betaine biosynthesis, choline => betaine **C,N**

M00020 Serine biosynthesis, glycerate-3P => serine **C,N**

M00033 Ectoine biosynthesis, aspartate => ectoine **C,N**

Cysteine and methionine metabolism

M00021 Cysteine biosynthesis, serine => cysteine **C,N**

M00338 Cysteine biosynthesis, homocysteine + serine => cysteine **C,N**

M00034 Methionine salvage pathway **C,N**

M00035 Methionine degradation **C,N**

M00017 Methionine biosynthesis, apartate => homoserine => methionine **C,N**

M00368 Ethylene biosynthesis, methionine => ethylene **C,N**

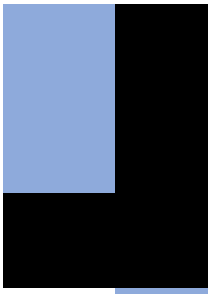
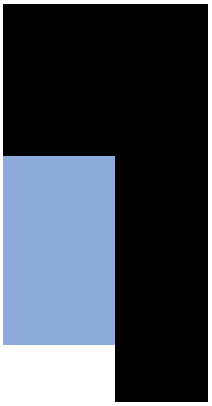
Branched-chain amino acid metabolism

M00019 Valine/isoleucine biosynthesis, pyruvate => valine / 2-oxobutanoate => isoleucine **C,N**

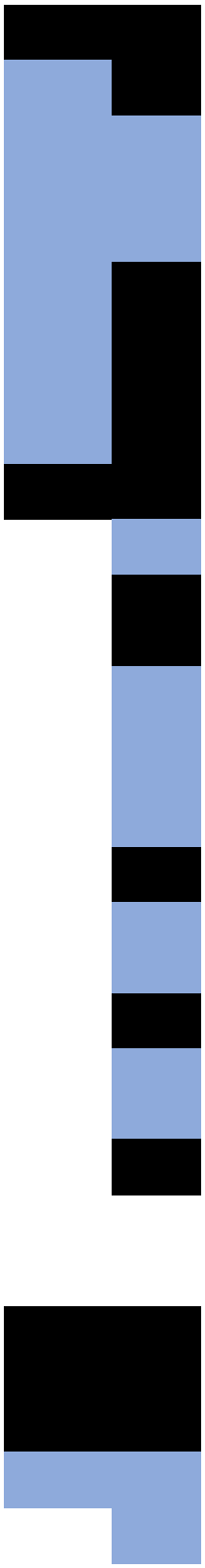
M00535 Isoleucine biosynthesis, pyruvate => 2-oxobutanoate **C,N**

M00570 Isoleucine biosynthesis, threonine => 2-oxobutanoate => isoleucine **C,N**

M00432 Leucine biosynthesis, 2-oxoisovalerate => 2-oxoisocaproate **C,N**

M00036 Leucine degradation, leucine => acetoacetate + acetyl-CoA		**C,N**
Lysine metabolism		
M00016 Lysine biosynthesis, succinyl-DAP pathway, aspartate => lysine		**C,N**
M00526 Lysine biosynthesis, DAP dehydrogenase pathway, aspartate => lysine		**C,N**
M00527 Lysine biosynthesis, DAP aminotransferase pathway, aspartate => lysine		**C,N**
M00433 Lysine biosynthesis, 2-oxoglutarate => 2-oxoadipate		**C,N**
Arginine and proline metabolism		
M00028 Ornithine biosynthesis, glutamate => ornithine		**C,N**
M00844 Arginine biosynthesis, ornithine => arginine		**C,N**
M00015 Proline biosynthesis, glutamate => proline		**C,N**
M00845 Arginine biosynthesis, glutamate => acetylcitrulline => arginine		**C,N**
M00047 Creatine pathway		**C,N**
Polyamine biosynthesis		
M00133 Polyamine biosynthesis, arginine => agmatine => putrescine => spermidine		**C,N**
M00134 Polyamine biosynthesis, arginine => ornithine => putrescine		**C,N**
M00136 GABA biosynthesis, prokaryotes, putrescine => GABA		**C,N**
Aromatic amino acid metabolism		
M00022 Shikimate pathway, phosphoenolpyruvate + erythrose-4P => chorismate		**C,N**
M00023 Tryptophan biosynthesis, chorismate => tryptophan		**C,N**
M00024 Phenylalanine biosynthesis, chorismate => phenylalanine		**C,N**
M00040 Tyrosine biosynthesis, prephanate => pretyrosine => tyrosine		**C,N**
M00025 Tyrosine biosynthesis, chorismate => tyrosine		**C,N**

M00042 Catecholamine biosynthesis, tyrosine => dopamine => noradrenaline => adrenaline	**C,N**
M00044 Tyrosine degradation, tyrosine => homogentisate	**C,N**
M00533 Homoprotocatechuate degradation, homoprotocatechuate => 2-oxohept-3-enedioate	**C,N**
M00545 Trans-cinnamate degradation, trans-cinnamate => acetyl-CoA	**C,N**
M00038 Tryptophan metabolism, tryptophan => kynurenine => 2-aminomuconate	**C,N**
Glycan metabolism	
Lipopolysaccharide metabolism	
M00064 ADP-L-glycero-D-manno-heptose biosynthesis	**C**
M00072 N-glycosylation by oligosaccharyltransferase	**C**
M00073 N-glycan precursor trimming	**C**
M00056 O-glycan biosynthesis, mucin type core	**C**
M00070 Glycosphingolipid biosynthesis, lacto-series, LacCer => Lc4Cer	**C**
M00068 Glycosphingolipid biosynthesis, globo-series, LacCer => Gb4Cer	**C**
M00060 KDO2-lipid A biosynthesis, Raetz pathway, LpxL-LpxM type	**C**
M00866 KDO2-lipid A biosynthesis, Raetz pathway, non-LpxL-LpxM type	**C**
M00063 CMP-KDO biosynthesis	**C**
M00064 ADP-L-glycero-D-manno-heptose biosynthesis	**C**
Metabolism of cofactors and vitamins	
Cofactor and vitamin metabolism	**C,P**
M00127 Thiamine biosynthesis, AIR => thiamine-P/thiamine-2P	**C,P**
M00125 Riboflavin biosynthesis, GTP => riboflavin/FMN/FAD	**C,P**
M00115 NAD biosynthesis, aspartate => NAD	**C,N**
M00119 Pantothenate biosynthesis, valine/L-aspartate => pantothenate	**C**



M00120 Coenzyme A biosynthesis, pantothenate => CoA	**C**
M00123 Biotin biosynthesis, pimeloyl-ACP/CoA => biotin	**C**
M00842 Tetrahydrobiopterin biosynthesis, GTP => BH4	**C**
M00843 L-threo-Tetrahydrobiopterin biosynthesis, GTP => L-threo-BH4	**C**
M00140 C1-unit interconversion, prokaryotes	**C**
M00121 Heme biosynthesis, plants and bacteria, glutamate => heme	**C, Fe**
M00846 Siroheme biosynthesis, glutamate => siroheme	**C,N,Fe**
M00112 Tocopherol/tocotorienol biosynthesis	**C**
M00124 Pyridoxal biosynthesis, erythrose-4P => pyridoxal-5P	**C,P**
M00572 Pimeloyl-ACP biosynthesis, BioC-BioH pathway, malonyl-ACP => pimeloyl-ACP	**C,P**
M00573 Biotin biosynthesis, BioI pathway, long-chain-acyl-ACP => pimeloyl-ACP => biotin	**C,P**
M00577 Biotin biosynthesis, BioW pathway, pimelate => pimeloyl-CoA => biotin	**C**
M00126 Tetrahydrofolate biosynthesis, GTP => THF	**C**
M00841 Tetrahydrofolate biosynthesis, mediated by PTPS, GTP => THF	**C**
M00122 Cobalamin biosynthesis, cobinamide => cobalamin	**C,Co**
M00117 Ubiquinone biosynthesis, prokaryotes, chorismate => ubiquinone	**C**
M00116 Menaquinone biosynthesis, chorismate => menaquinol	**C**

Biosynthesis of terpenoids and polyketides

Terpenoid backbone biosynthesis

M00096 C5 isoprenoid biosynthesis, non-mevalonate pathway	**C**
M00364 C10-C20 isoprenoid biosynthesis, bacteria	**C**
M00365 C10-C20 isoprenoid biosynthesis, archaea	**C**
M00095 C5 isoprenoid biosynthesis, mevalonate pathway	**C**

Xenobiotics biodegradation

Aromatics degradation

M00538 Toluene degradation, toluene => benzoate **C**

M00537 Xylene degradation, xylene => methylbenzoate **C**

M00551 Benzoate degradation, benzoate => catechol / methylbenzoate => methylcatechol lete) **C**

M00637 Anthranilate degradation, anthranilate => catechol **C**

M00568 Catechol ortho-cleavage, catechol => 3-oxoadipate **C**

M00569 Catechol meta-cleavage, catechol => acetyl-CoA / 4-methylcatechol => propanoyl-CoA **C**

M00540 Benzoate degradation, cyclohexanecarboxylic acid =>pimeloyl-CoA **C**

M00638 Salicylate degradation, salicylate => gentisate **C**

M00623 Phthalate degradation, phthalate => protocatechuate **C**

Drug resistance

M00627 beta-Lactam resistance, Bla system

M00745 Imipenem resistance, repression of porin OprD (13)

M00651 Vancomycin resistance, D-Ala-D-Lac type

M00726 Cationic antimicrobial peptide (CAMP) resistance, lysyl-phosphatidylglycerol (L-PG) synthase MprF

M00744 Cationic antimicrobial peptide (CAMP) resistance, protease PgtE

M00718 Multidrug resistance, efflux pump MexAB-OprM

M00642 Multidrug resistance, efflux pump MexJK-OprM

M00643 Multidrug resistance, efflux pump MexXY-OprM

M00769 Multidrug resistance, efflux pump MexPQ-OpmE

M00649 Multidrug resistance, efflux pump AdeABC

M00696 Multidrug resistance, efflux pump AcrEF-TolC

M00697 Multidrug resistance, efflux pump MdtEF-TolC

M00698 Multidrug resistance, efflux pump BpeEF-OprC

M00700 Multidrug resistance, efflux pump AbcA

M00714 Multidrug resistance, efflux pump QacA

Appendix 2

Supplemental Material to Chapter 2:

Cyanobacterial bloom succession effects on bacteria community composition over multiple years

Katherine V. Cook, Jessica E. Beyer, and K. David Hambright

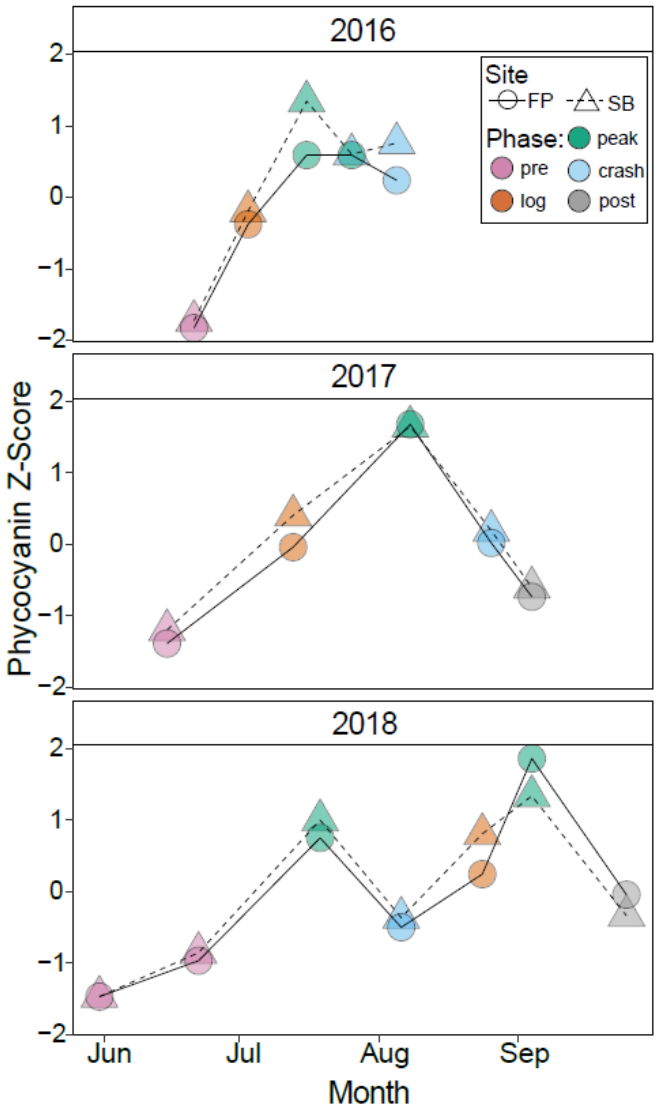


Figure S1. Bloom phases from the three sampling years. Phases were assigned by assessing PCY z-score differences between samples within the same year, paired with knowledge of the system and sites. (A) shows the 2016 bloom, (B) 2017, and (C) 2018, where more samples were sequenced because of the two bloom peaks. Fisherman's Point site is indicated by the solid line and Sailboat is indicated by the dashed line.

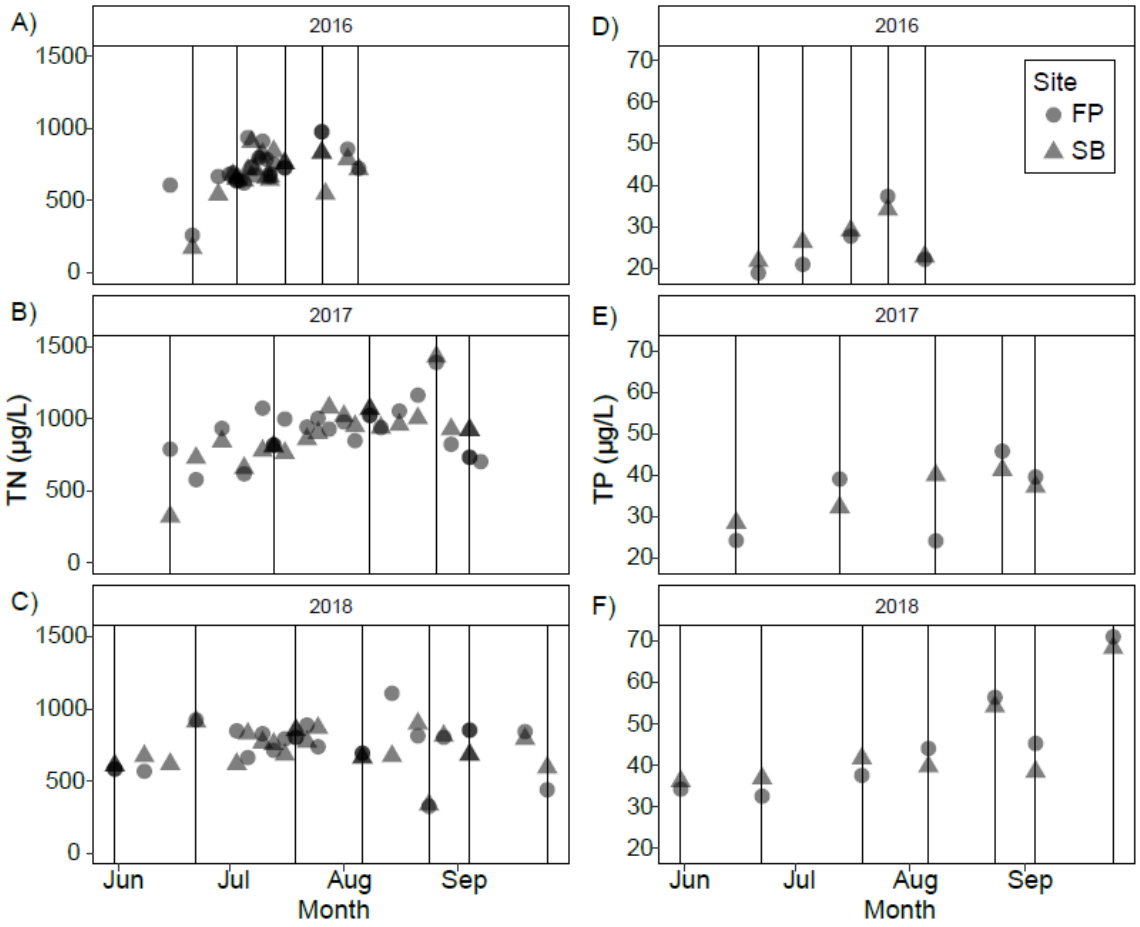


Figure S2. Nutrient concentrations across three years of sampling: 2016 total nitrogen (TN) (A), 2017 TN (B), 2018 TN (C), 2016 total phosphorus (TP) (D), 2017 TP (E), and 2018 TP (F). The sites are indicated by different shapes, circle for Fisherman's Point and triangle for Sailboat. Darker areas indicate increased overlap of points. Vertical lines indicate bloom phase samples used for sequencing and subsequent data analysis. There was unequal sample analysis between TN and TP and not all sampling days were analyzed for TN, but all sequenced samples were analyzed for both nutrients.

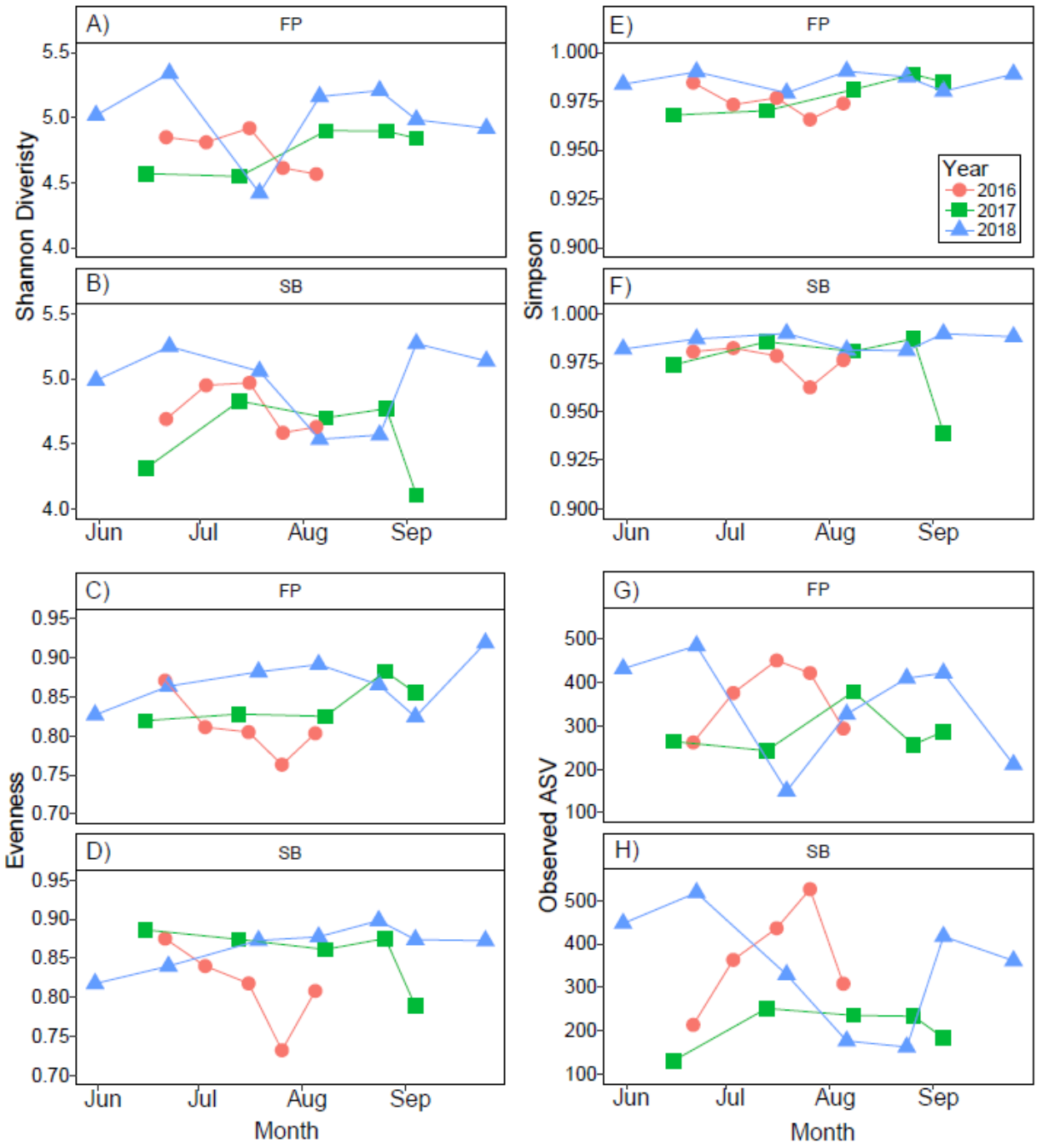


Figure S3. Alpha diversity metrics for each community in both sites for all three years. The sites are abbreviated, Fisherman's Point - FP and Sailboat - SB. The years are distinguished by different colors and shapes, 2016 is red and a circle, 2017 is green and a square, and 2018 is blue and a triangle. The panels are (A-B) Shannon diversity, (C-D) Evenness, (E-F) Simpson's D (1-D), and (G-H) Observed ASV. Higher numbers of Shannon Diversity indicates higher diversity.

Evenness ranges from 0 to 1 and higher numbers indicating higher evenness (many species with similar abundances). Simpson's D (1-D) ranges from 0 to 1 values closer to 1 indicates high diversity with an even proportion of abundance.

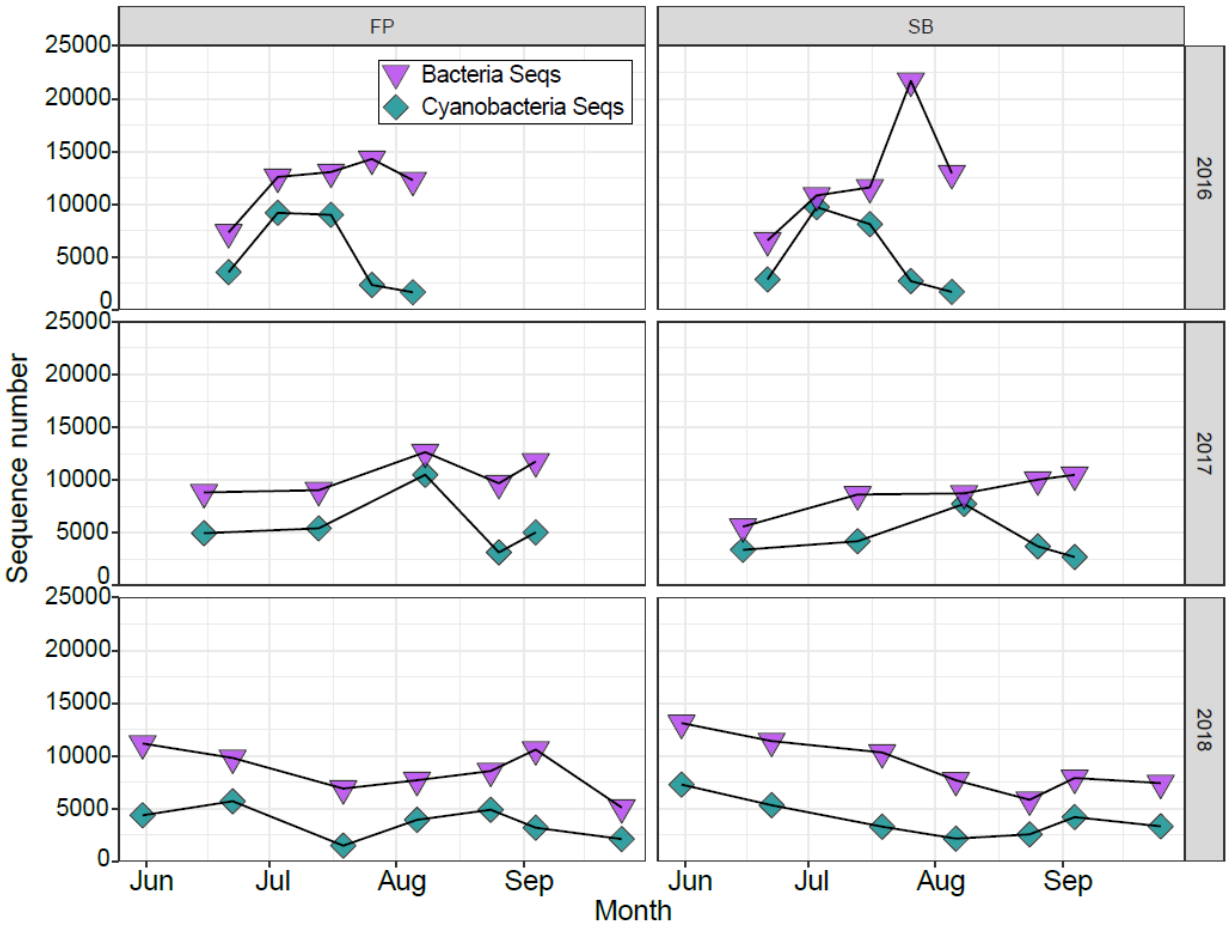


Fig S4. Number of bacteria (purple upside-down triangle) and cyanobacteria (cyan diamond) sequences in each sample. The first column is Fisherman's Point (FP) and the second column is Sailboat (SB). The first row is year 2016, the second row is year 2017, and the third row is year 2018. Position corresponds with sampling date.

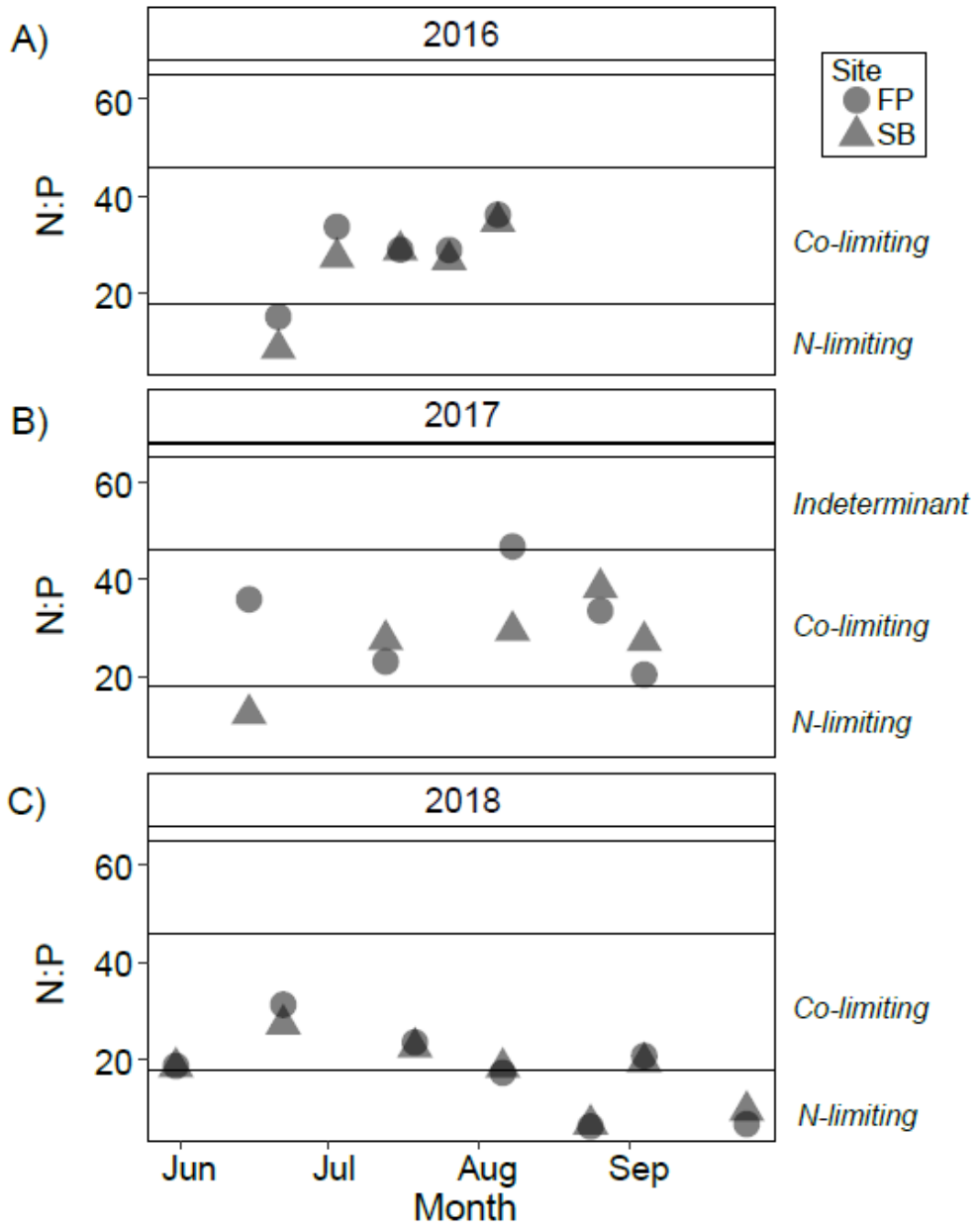


Fig S5. Nitrogen to phosphorus molar ratios (N:P) on Lake Thunderbird at Fisherman's Point (FP, circles) and Sailboat (SB, triangles) over three years. Below the horizontal line at 18 indicates nitrogen limiting (n-limiting) conditions and between 18 and 45 indicates both nitrogen and phosphorus limitation (co-limiting), 45 - 65 is indeterminant, and above 65 is phosphorus limitation (never reached in this study).

Appendix 3

Supplemental Material to Chapter 3:

Ground-based remote sensing provides alternative to satellites for monitoring cyanobacterial blooms in small lakes

Katherine V. Cook, Jessica E. Beyer, Xiangming Xiao, and K. David Hambright

---

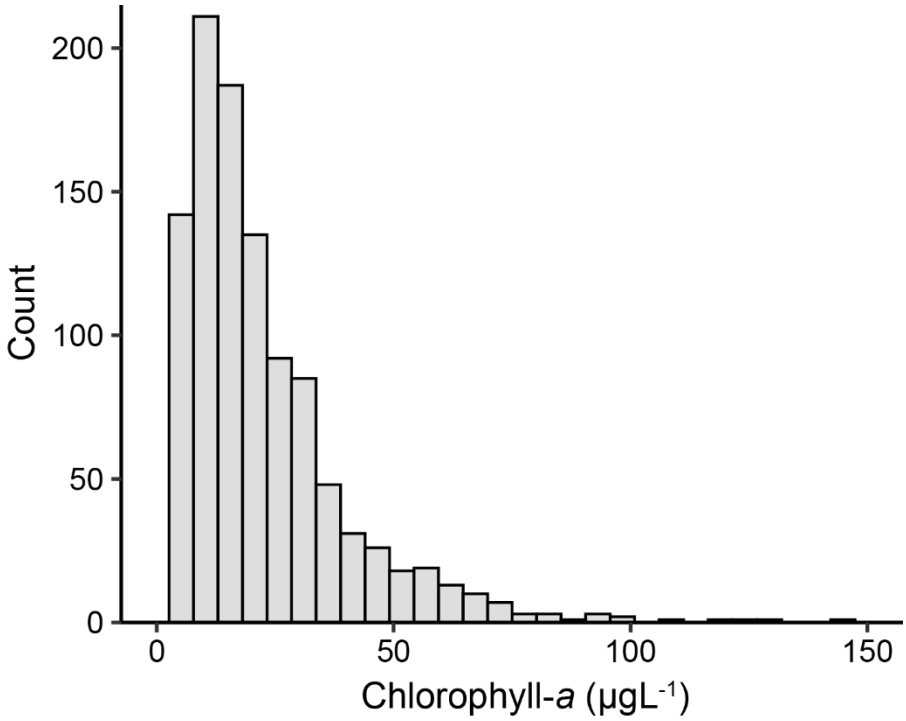


Figure S1. Distribution of chlorophyll-*a* in the full satellite dataset (n = 1160).

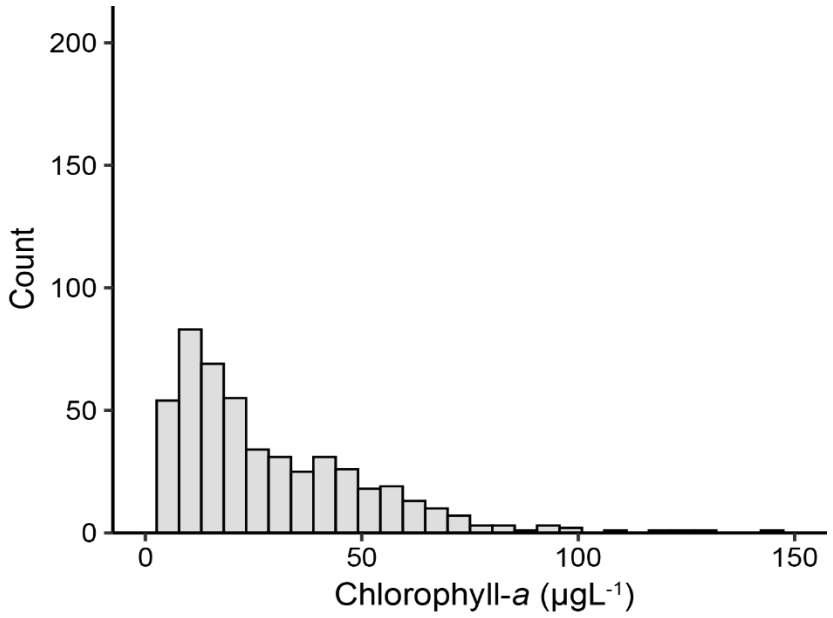


Figure S2. Distribution of chlorophyll-*a* in the re-sampled dataset (n = 511).

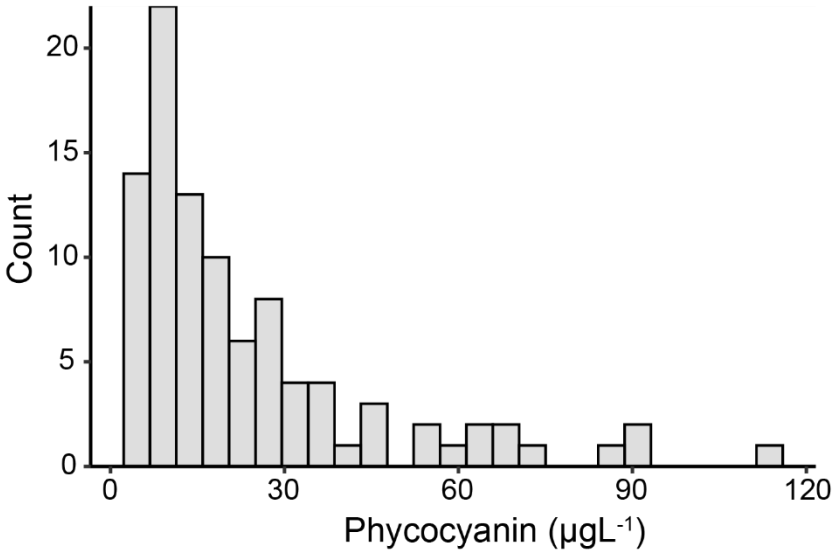


Figure S3. Distribution of phycocyanin in the full satellite dataset (n = 97).

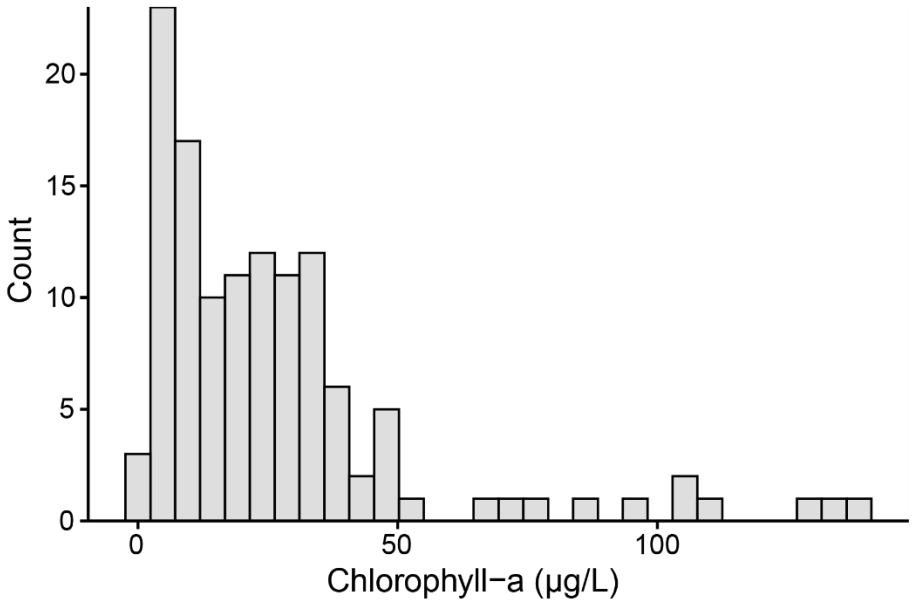


Figure S4. Distribution of chlorophyll-*a* in the ground-based dataset (n = 124).

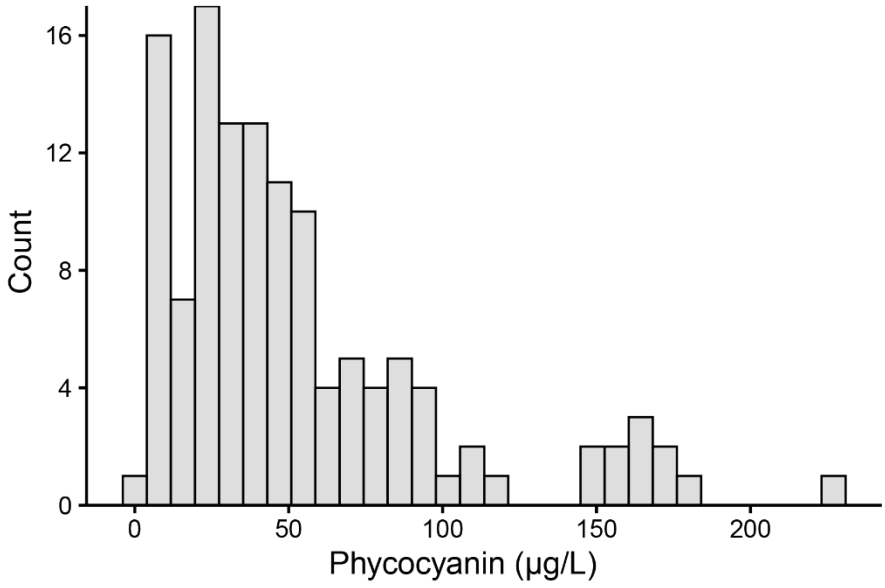


Figure S5. Distribution of phycocyanin in the ground-based dataset (n = 125).

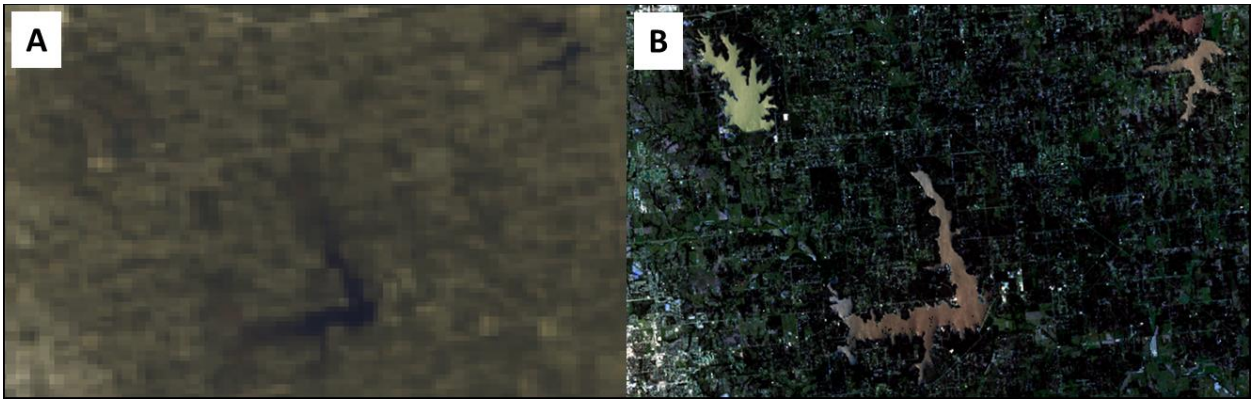


Figure S6. Comparison of Sentinel-3a and Landsat 8 images to demonstrate how the difference in pixel size impacts lake resolvability. Lake Thunderbird is pictured here at the center of each image, the Shawnee Twin Lakes in the upper right-hand corner, and Lake Stanley Draper to the upper left-hand corner of the images. (A) shows a Sentinel-3b (pixel size 300 x 300 m) false color composite of bands 1, 2, and 3 built using Sentinels Application Platform (SNAP) from the European Space Agency (ESA) and (B) shows a Landsat-8 (pixel size 30 x 30 m) false color composite of bands 2, 3, and 4 built in QGIS.

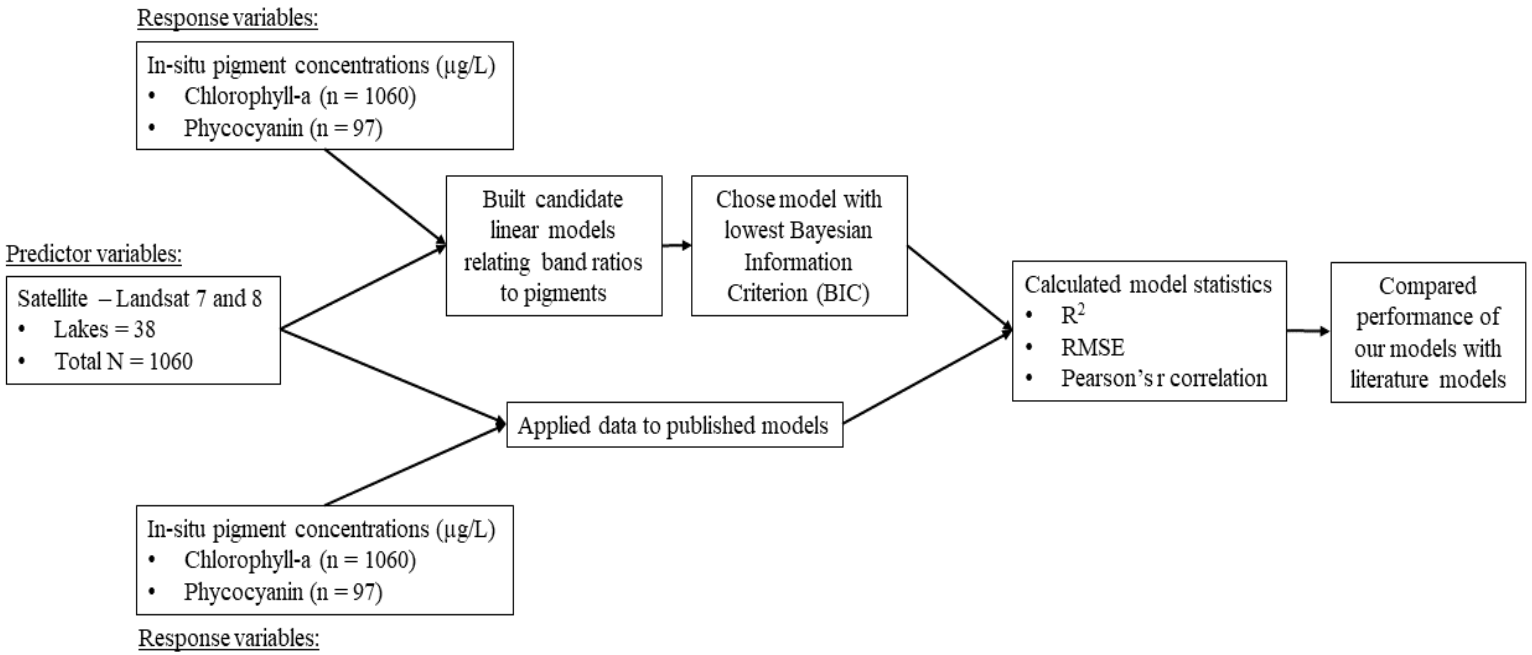


Figure S7. Schematic workflow diagram for satellite methods.

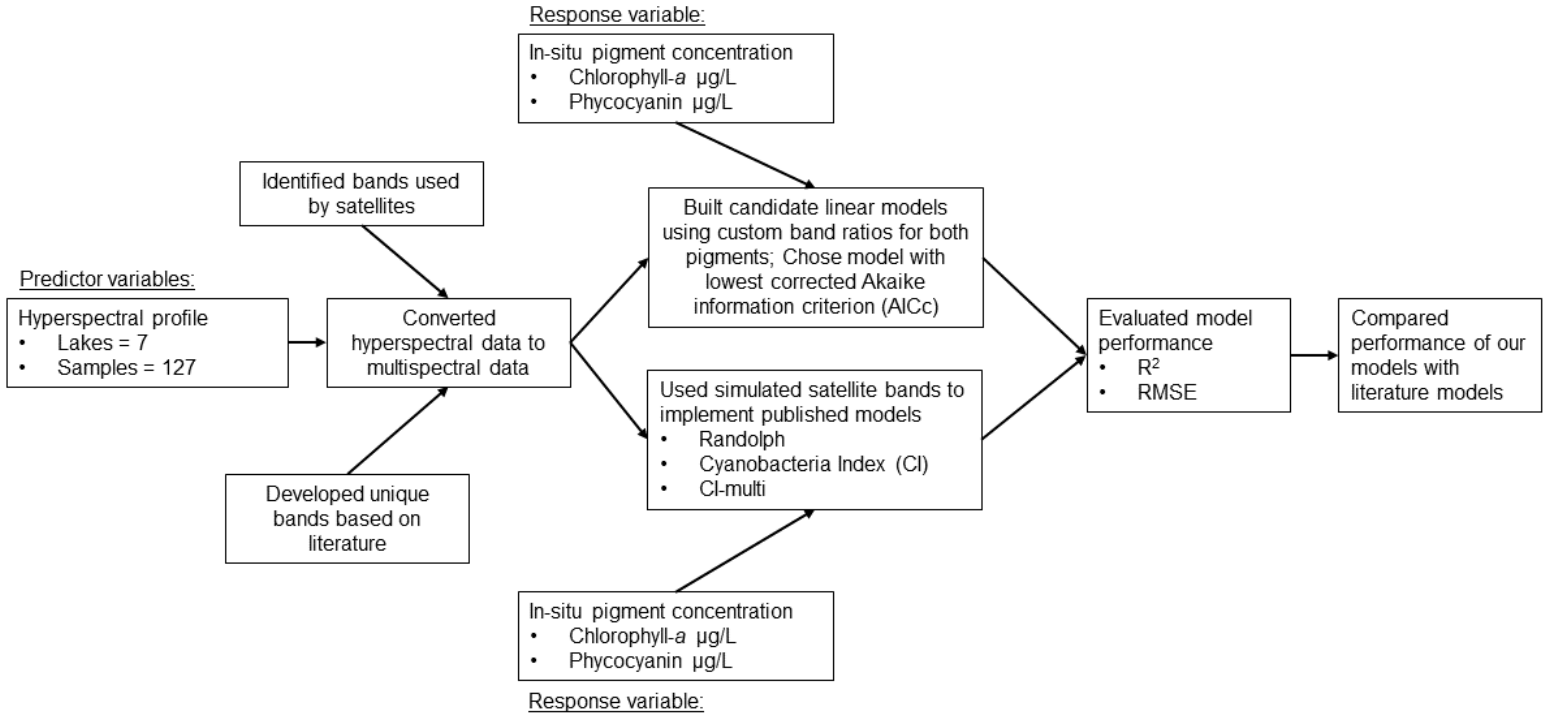


Figure S8. Schematic workflow diagram for ground-based methods.

Table S1. Site information for the ground-truthed satellite chlorophyll-*a* data. Sampling organizations included Oklahoma Water Resources Board (OWRB), Plankton Ecology and Limnology Lab (PELL), and Grand River Dam Authority (GRDA).

Lake	Site	Number of samples	Sampling organization	Latitude	Longitude	Lake area (hectare)
Altus	Site 2	3	OWRB	34.90247	-99.2937	2,600
Altus	Site 3	2	OWRB	34.92733	-99.3130	
Altus	Surface	2	OWRB	34.88722	-99.2949	
Arcadia	Site 2	5	OWRB	35.63511	-97.3714	737
Arcadia	Site 3	5	OWRB	35.64489	-97.3853	
Arcadia	Site 4	5	OWRB	35.62770	-97.3939	
Arcadia	Site 5	5	OWRB	35.61594	-97.4059	
Atoka	Site 2	6	OWRB	34.47904	-96.0908	2,307
Atoka	Site 3	5	OWRB	34.50523	-96.0754	
Atoka	Site 4	3	OWRB	34.54017	-96.0462	
Atoka	Surface	3	OWRB	34.44624	-96.0872	
Birch	77	1	OWRB	36.51479	-96.1912	460
Canton	Site 2	2	OWRB	36.12572	-98.6014	3,201
Canton	Site 3	3	OWRB	36.14255	-98.6339	
Canton	Site 4	1	OWRB	36.14261	-98.6549	
Canton	Site 5	1	OWRB	36.09548	-98.5867	
Canton	Surface	3	OWRB	36.09337	-98.5883	
Chickasha	Site 2	4	OWRB	35.14737	-98.1502	837
Chickasha	Site 3	4	OWRB	35.14995	-98.1307	
Chickasha	Site 4	1	OWRB	35.13559	-98.1402	
Chickasha	Site 5	1	OWRB	35.14373	-98.1368	
Chickasha	Surface	4	OWRB	35.13199	-98.1329	
Claremore	Site 2	4	OWRB	36.34031	-95.5725	190
Claremore	Site 3	4	OWRB	36.34092	-95.5599	
Claremore	Site 4	2	OWRB	36.33351	-95.5786	
Claremore	Surface	4	OWRB	36.32514	-95.5796	
Clinton	Site 3	1	OWRB	35.42906	-99.2231	136
Copan	Site 2	4	OWRB	36.91109	-95.9564	1,963
Copan	Site 3	4	OWRB	36.92622	-95.9528	
Copan	Site 4	4	OWRB	36.94839	-95.9401	
Copan	Site 5	4	OWRB	36.93934	-95.9559	
Copan	Surface	4	OWRB	36.88641	-95.9679	

Crowder	Site 2	11	OWRB	35.39980	-98.7074	64
Cushing	80	2	NLA	36.00416	-96.879	170
Dave Boyer	73	2	NLA	34.37438	-98.3367	49
El Reno	2	5	PELL	35.52391	-97.9897	69
El Reno	3	5	PELL	35.51955	-97.9962	
Ellsworth	Site 2	5	OWRB	34.82185	-98.356	2,266
Ellsworth	Site 4	4	OWRB	34.81763	-98.3397	
Ellsworth	Site 5	3	OWRB	34.84256	-98.3584	
Eufaula	Site 11	6	OWRB	35.22886	-95.6338	42,695
Eufaula	Site 12	5	OWRB	35.20047	-95.5938	
Eufaula	Site 13	5	OWRB	35.16440	-95.5995	
Eufaula	Site 14	5	OWRB	35.10144	-95.6472	
Eufaula	Site 15	4	OWRB	35.05004	-95.6711	
Eufaula	Site 16	4	OWRB	35.01996	-95.6023	
Eufaula	Site 17	4	OWRB	34.97471	-95.6303	
Eufaula	Site 2	6	OWRB	35.42767	-95.6001	
Eufaula	Site 3	5	OWRB	35.38221	-95.6300	
Eufaula	Site 4	6	OWRB	35.30045	-95.5540	
Eufaula	Site 5	4	OWRB	35.28483	-95.5147	
Eufaula	Site 6	5	OWRB	35.30704	-95.4376	
Eufaula	Site 8	6	OWRB	35.23392	-95.4999	
Eufaula	Site 9	6	OWRB	35.22526	-95.5963	
Eufaula	Surface	4	OWRB	35.45436	-95.6129	
Fort Cobb	Site 1	4	OWRB	35.16215	-98.4568	1,619
Fort Cobb	Site 2	5	OWRB	35.18004	-98.4623	
Fort Cobb	Site 3	5	OWRB	35.18973	-98.4734	
Fort Cobb	Site 4	3	OWRB	35.21948	-98.5095	
Fort Cobb	Site 5	4	OWRB	35.20804	-98.4911	
Fort Cobb	Site 6	1	OWRB	35.21919	-98.4804	
Foss	Site 2	4	OWRB	35.56204	-99.2101	3,561
Foss	Site 3	5	OWRB	35.57753	-99.2352	
Foss	Site 4	5	OWRB	35.60198	-99.2327	
Foss	Site 5	5	OWRB	35.61109	-99.2619	
Foss	Surface	4	OWRB	35.53936	-99.1890	
Grand	Site 10	3	OWRB	36.66528	-94.7703	16,908
Grand	Site 11	3	OWRB	36.65250	-94.7183	
Grand	Site 12	3	OWRB	36.69889	-94.7431	
Grand	Site 13	3	OWRB	36.73333	-94.7739	
Grand	Site 2	2	OWRB	36.51028	-94.9650	
Grand	Site 3	2	OWRB	36.54417	-94.9303	
Grand	Site 4	2	OWRB	36.60361	-94.9039	

Grand	Site 5	3	OWRB	36.56361	-94.8606	
Grand	Site 6	1	OWRB	36.56306	-94.7692	
Grand	Site 7	3	OWRB	36.57194	-94.8331	
Grand	Site 8	3	OWRB	36.62278	-94.8433	
Grand	Site 9	3	OWRB	36.63750	-94.8014	
Grand	Surface	3	OWRB	36.47472	-95.0347	
Grand	12	5	GRDA	36.64998	-94.7084	
Grand	13	6	GRDA	36.68269	-94.7728	
Grand	14	5	GRDA	36.57437	-94.7899	
Grand	15	5	GRDA	36.62237	-94.9079	
Grand	18	1	GRDA	36.49758	-95.0109	
Grand	2	1	GRDA	36.62441	-94.9007	
Grand	21	1	GRDA	36.50138	-94.9242	
Grand	22	6	GRDA	36.55447	-94.8449	
Grand	26	5	GRDA	36.56396	-94.9128	
Grand	27	1	GRDA	36.53917	-94.8389	
Grand	28	1	GRDA	36.53807	-94.8352	
Grand	29	5	GRDA	36.54386	-94.8436	
Grand	7	6	GRDA	36.49768	-94.9185	
Grand	8	1	GRDA	36.50451	-94.9646	
Greenleaf	Site 2	4	OWRB	35.63017	-95.1600	229
Greenleaf	Site 3	4	OWRB	35.64416	-95.1522	
Greenleaf	Site 4	2	OWRB	35.64852	-95.1408	
Greenleaf	Site 5	2	OWRB	35.62183	-95.1605	
Greenleaf	Surface	3	OWRB	35.61719	-95.1664	
Hefner	Site 2	3	OWRB	35.56324	-97.6041	1,000
Hefner	Site 3	3	OWRB	35.56239	-97.5827	
Hefner	Site 4	1	OWRB	35.55528	-97.5929	
Hefner	Site 5	1	OWRB	35.57542	-97.5896	
Hefner	Surface	2	OWRB	35.58102	-97.5974	
Hudson	34	2	GRDA	36.37373	-95.1225	4,856
Hudson	35	2	GRDA	36.30724	-95.1813	
Hudson	87	1	OWRB	36.82300	-96.0476	
Hulah	85	1	OWRB	36.93107	-96.1030	1,445
Jean Neustadt	90	1	OWRB	34.28478	-97.1710	187
Kaw	Site 2	5	OWRB	36.74621	-96.8831	6,879
Kaw	Site 3	4	OWRB	36.76728	-96.8269	
Kaw	Site 4	3	OWRB	36.79969	-96.8291	
Kaw	Site 5	3	OWRB	36.79038	-96.9080	
Kaw	Surface	2	OWRB	36.70118	-96.9242	
Keystone	Site 10	4	OWRB	36.19266	-96.3167	10,523

Keystone	Site 2	3	OWRB	36.19016	-96.2537	
Keystone	Site 3	4	OWRB	36.23228	-96.3001	
Keystone	Site 4	4	OWRB	36.23771	-96.3551	
Keystone	Site 6	4	OWRB	36.16468	-96.2934	
Keystone	Site 8	4	OWRB	36.13969	-96.3285	
Keystone	Site 9	4	OWRB	36.16656	-96.3126	
Keystone	Surface	3	OWRB	36.14726	-96.2573	
Keystone	89	1	OWRB	36.15061	-96.4397	
New Spiro	Site 2	3	OWRB	35.19968	-94.6198	83
Okemah	Site 2	2	OWRB	35.52537	-96.3203	308
Okemah	Site 3	2	OWRB	35.52375	-96.3323	
Okemah	Site 4	1	OWRB	35.50891	-96.3231	
Overholser	Site 2	4	OWRB	35.49974	-97.6766	640
Overholser	Site 3	3	OWRB	35.50848	-97.6700	
Overholser	Site 5	1	OWRB	35.50661	-97.6777	
Overholser	Surface	4	OWRB	35.48716	-97.6682	
Overholser	75	1	OWRB	35.49774	-97.6793	
Perry	Site 2	4	OWRB	36.24243	-97.3409	227
Perry	Site 3	3	OWRB	36.23827	-97.3496	
Perry	Site 5	2	OWRB	36.24680	-97.3398	
Ponca	Site 2	2	OWRB	36.73613	-97.0343	326
Ponca	Site 4	2	OWRB	36.72930	-97.0294	
Ponca	Site 5	2	OWRB	36.71885	-97.0167	
RC Longmire	Site 2	3	OWRB	34.75152	-97.0511	301
RC Longmire	Site 4	2	OWRB	34.75060	-97.0447	
RC Longmire	Site 5	1	OWRB	34.75533	-97.0521	
Rocky (Hobart)	Site 2	2	OWRB	35.17303	-99.0768	138
Rocky (Hobart)	Site 3	3	OWRB	35.18312	-99.0738	
Rocky (Hobart)	Site 4	2	OWRB	35.17694	-99.0788	
Taylor (Marlow)	Site 2	2	OWRB	34.75048	-97.9278	92
Tenkiller Ferry	Site 2	7	OWRB	35.67443	-94.9764	5,221
Tenkiller Ferry	Site 3	5	OWRB	35.73905	-94.9543	
Tenkiller Ferry	Site 4	3	OWRB	35.75542	-94.9051	
Tenkiller Ferry	Site 6	2	OWRB	35.76634	-94.8872	
Tenkiller Ferry	Site 7	6	OWRB	35.63938	-95.0146	
Tenkiller Ferry	Surface	2	OWRB	35.60002	-95.0446	
Tenkiller Ferry	91	1	OWRB	35.75412	-94.9149	
Texoma	Site 10	3	OWRB	33.78593	-96.7962	35,613
Texoma	Site 11	4	OWRB	33.86820	-96.8383	
Texoma	Site 12	4	OWRB	33.89428	-96.8888	
Texoma	Site 2	3	OWRB	33.89302	-96.6113	

Texoma	Site 3	3	OWRB	33.95756	-96.5871	
Texoma	Site 4	3	OWRB	34.00762	-96.6299	
Texoma	Site 6	4	OWRB	33.85604	-96.6915	
Texoma	Site 7	4	OWRB	33.82814	-96.7376	
Texoma	Site 8	4	OWRB	33.84694	-96.7801	
Texoma	Site 9	4	OWRB	33.81989	-96.8054	
Texoma	Surface	4	OWRB	33.82997	-96.5768	
Texoma	Buncombe North 5	1	PELL	33.87486	-96.8073	
Texoma	Buncombe Pelagic 2	30	PELL	33.87125	-96.8074	
Texoma	Buncombe South 1	1	PELL	33.86158	-96.8069	
Texoma	Dam North	1	PELL	33.83689	-96.5892	
Texoma	Dam Pelagic 4	27	PELL	33.82319	-96.5902	
Texoma	Islands Pelagic 3	30	PELL	33.82875	-96.7304	
Texoma	Red River Pelagic 1	30	PELL	33.89742	-96.8874	
Texoma	Washita Pelagic 5	29	PELL	33.96403	-96.5769	
Texoma	Washita South 1	1	PELL	33.89358	-96.5808	
Thunderbird	1	42	OWRB	35.22333	-97.2208	2,456
Thunderbird	2	56	OWRB	35.23889	-97.2289	
Thunderbird	3	54	OWRB	35.26222	-97.2389	
Thunderbird	4	56	OWRB	35.22444	-97.2508	
Thunderbird	5	54	OWRB	35.22028	-97.2906	
Thunderbird	7	23	OWRB	35.20306	-97.2581	
Thunderbird	8	31	OWRB	35.28641	-97.2449	
Thunderbird	Fisherman's point	4	PELL	35.22862	-97.2460	
Thunderbird	North Sentinel	4	PELL	35.23219	-97.3075	
Webbers Falls	Site 2	4	OWRB	35.60167	-95.1817	4,694
Webbers Falls	Site 4	2	OWRB	35.63056	-95.2717	
Webbers Falls	Site 6	4	OWRB	35.70000	-95.2314	
Webbers Falls	Surface	3	OWRB	35.55472	-95.1706	
WRHoloway	Site 2	6	OWRB	36.24564	-95.0996	318
WRHoloway	Site 3	9	OWRB	36.25695	-95.0868	
WRHoloway	Site 4	3	OWRB	36.25100	-95.1021	
WRHoloway	Site 5	1	OWRB	36.25209	-95.0931	
WRHoloway	Surface	10	OWRB	36.25634	-95.1023	

WRHoloway	36	2	GRDA	36.25576	-95.1006	
WRHoloway	38	1	GRDA	36.25562	-95.0900	

Table S2. Site information for the satellite ground-truthed phycocyanin samples. Sampling organization included the Plankton Ecology and Limnology Lab (PELL).

Lake	Site	Number of samples	Sampling organization	Latitude	Longitude
Texoma	Buncombe North 5	1	PELL	33.87486	-96.8073
Texoma	Buncombe Pelagic 2	17	PELL	33.87125	-96.8074
Texoma	Buncombe South 1	1	PELL	33.86158	-96.8069
Texoma	Dam North	1	PELL	33.83689	-96.5892
Texoma	Dam Pelagic 4	17	PELL	33.82319	-96.5902
Texoma	Islands Pelagic 3	17	PELL	33.82875	-96.7304
Texoma	Red River Pelagic 1	17	PELL	33.89742	-96.8874
Texoma	Washita Pelagic 5	17	PELL	33.96403	-96.5769
Texoma	Washita South 1	1	PELL	33.89358	-96.5808
Thunderbird	Fisherman's Point	4	PELL	35.22862	-97.2460
Thunderbird	North Sentinel	4	PELL	35.23219	-97.3075

Table S3. Ground-based site information. Sampling organization included the Plankton Ecology and Limnology Lab (PELL).

Lake	Site	Number of samples	Sampling organization	Latitude	Longitude
Ellsworth	Dam	1	PELL	34.79550	-98.36640
El Reno	1	1	PELL	35.52718	-97.98684
El Reno	2	1	PELL	35.52373	-97.98925
El Reno	3	1	PELL	35.51960	-97.99631
El Reno	4	1	PELL	35.52640	-97.98934
El Reno	5	1	PELL	35.52110	-97.99271
Grand	Drip	4	PELL	36.49969	-94.95614
Grand	Drown	4	PELL	36.49842	-94.91957
Grand	Duck	5	PELL	36.53644	-94.97219
Grand	Grand	5	PELL	36.68269	-94.77286
Grand	Honey	5	PELL	36.57511	-94.78772
Grand	Horse	5	PELL	36.62122	-94.90856
Grand	IS1	5	PELL	36.49250	-95.04489
Grand	IS3	5	PELL	36.50858	-94.95539
Grand	Sail	5	PELL	36.64175	-94.81478
Grand	Tree	5	PELL	36.56339	-94.91283
Grand	Wood	5	PELL	36.53644	-94.82236
Overholser	1	1	PELL	35.48720	-97.66820
Overholser	2	1	PELL	35.49970	-97.67660
Overholser	4	1	PELL	35.49230	-97.67560
Rocky	Dam	1	PELL	35.16720	-99.07430
Texoma	Buncombe North 4	1	PELL	33.87317	-96.80725
Texoma	Buncombe North 5	3	PELL	33.87472	-96.80722
Texoma	Buncombe Pelagic 2	4	PELL	33.87125	-96.80744
Texoma	Buncombe South 1	2	PELL	33.86158	-96.80692
Texoma	Buncombe South 2	1	PELL	33.86572	-96.80764
Texoma	Dam North	3	PELL	33.83689	-96.58919
Texoma	Dam Pelagic 4	4	PELL	33.82319	-96.59022
Texoma	Islands Pelagic 3	4	PELL	33.82875	-96.73044
Texoma	Red River Pelagic 1	3	PELL	33.89481	-96.89172
Texoma	Washita Pelagic 5	4	PELL	33.96403	-96.57692

Texoma	Washita South 1	3	PELL	33.87553	-96.61975
Texoma	Washita South 2	3	PELL	33.92900	-96.57131
Thunderbird	1	2	PELL	35.22275	-97.22237
Thunderbird	2	2	PELL	35.23855	-97.22916
Thunderbird	3	2	PELL	35.26260	-97.23883
Thunderbird	4	2	PELL	35.22381	-97.25139
Thunderbird	5	1	PELL	35.22016	-97.28863
Thunderbird	6	1	PELL	35.23065	-97.30523
Thunderbird	8	2	PELL	35.28664	-97.24446
Thunderbird	11	2	PELL	35.21296	-97.30335
Thunderbird	13	1	PELL	-	-
Thunderbird	Fisherman's Point	4	PELL	35.22909	-97.24615
Thunderbird	North Sentinel	4	PELL	35.23194	-97.30929
Thunderbird	Sailboat	4	PELL	35.23037	-97.23629

Table S4. Performance of satellite algorithms on the resampled data set (N = 511) for chlorophyll-*a* estimation on Oklahoma lakes was evaluated using the adjusted R^2 from the given linear model, the root mean square error (RMSE), and Pearson's r correlation coefficient. *The RMSE for the resampled model should not be directly compared to the other models because the dataset used to build the model was different.

	Algorithm name	**Adjusted R^2**	**Model P-value**	**RMSE**	**Pearson's r**	**Pearson's r P-value**
	Chlorophyll-*a* algorithms					
1	NDCI (Normalized Difference Chlorophyll Index)	-0.0013	0.57	22.26	0.025	0.57
2	SABI (Surface Algal Bloom Index)	-0.0014	0.58	22.26	0.024	0.58
3	FLHB (Fluorescence Line Height algorithm blue)	-0.0014	0.59	22.26	0.024	0.59
4	2BDA (two-band algorithm)	-0.0007	0.42	22.26	0.036	0.42
5	3BDA (three-band algorithm)	-0.0007	0.43	22.26	0.035	0.43
6	KIVU (3BDA – like)	-0.0009	0.46	22.26	0.029	0.46
7	this paper (resampled dataset)	0.025	<0.001	21.95	0.170	<0.001

Equation for the ground-based chlorophyll-*a* model. For band information see Table 2.

CHLA = -222.71 – 100.43(Band1:Band3) + 107.2(Band1:Band5) – 34.52(Band2:Band3) + 564.61(Band3:Band4) – 856.16(Band3:Band5) + 564.63(Band4:Band5) (eq. S1)

Equation for the ground-based phycocyanin model. For band information see Table 2.

PCY = -1311.71 – 161.24 (Band1:Band2) + 145.35 (Band1:Band5) + 568.31 (Band2:Band3) – 836.64 (Band2:Band4) + 2074.41 (Band3:Band4) – 1370.80 (Band3:Band5) + 948.89 (Band4:Band5) (eq. S2)

R code for processing Landsat images in the R environment:

```
require(raster)
require(rgdal)
require(dplyr)
require(sp)

#set your working directory, all of your image folders should be in this folder
your_working_directory <- "K:/Images for qaqc - 10Nov20/Thunderbird8"

setwd(your_working_directory)

file.names <- list.files(your_working_directory)

for (x in 1:length(file.names)){
  print(x)
  holder <- paste0(your_working_directory, file.names[x], "/")
  file.names[x] <- holder
}

field_points <- read.csv(file="K:/Images for qaqc - 10Nov20/Thunderbird8.csv") #csv has to be set up with longitude in a column before latitude, can have as many lake site combinations as needed

#field_points2 <- select(field_points, -Lake)
field_points2 <- project(as.matrix(field_points[3:4]), proj="+proj=utm +zone=14 ellps=WGS84") ##use this if you need to change the projection of the field points, all of the images should be in the same projection for each field point file ##for field_points[x:y] put the longitude (x) and latitude (y) columns


#calculate the extent of the image you need to look at based on field points
e <- extent(min(field_points2[,1])-1000, max(field_points2[,1])+1000,
        min(field_points2[,2])-1000,max(field_points2[,2])+1000)
```

```
for(j in 1:length(file.names)){

  bands = list.files(path = paste(file.names[j]), pattern = "sr_band[1-7].tif$", recursive = TRUE)
##make sure this is correct

  for(i in 1:6){
    r=raster(paste(paste(file.names[j]),bands[i],sep=''))
    r <- crop(r, e)
    r <- focal(r, w=matrix(1/9, nrow=3, ncol=3), na.rm=TRUE)
    names(r) <- paste("layer",i)
    if(i==1){
      b=r
    }
    else {
      b=addLayer(b,r)
    }
    print(i)
  }

  #extract points of interest (these represent the average of the 3x3 neighborhood of the point of
interest)
  tempDF3x3 <- as.data.frame(extract(b,field_points2))

  #add info about site
  tempDF3x3$site <- c(seq(1,1))

  #add filename to df
  tempDF3x3$filename <- c(rep(file.names[j],1)) #is this the number of sites?

  #rename columns that hold the band data
  names(tempDF3x3)[1:6] <- c("band_1", "band_2", "band_3", "band_4", "band_5", "band_7")

if(j==1){
    masterDF <- tempDF3x3
  }
  else{
    masterDF <- bind_rows(masterDF, tempDF3x3)
  }
}

#save the surface reflectance to a csv
write.csv(masterDF, file='./Landsat_sr.csv')
```