Computational Ecology & Open Science: Tools to Help Manage Cyanobacteria in Lakes

Betty J. Kreakie, Jeffrey W. Hollister, Farnaz Nojavan, W. Bryan Milstead, and Lahne Mattas-Curry

We are a small group of computational ecologists tucked away in the U.S. EPA's Office of Research and Development lab on the coast of Narragansett Bay, Rhode Island. Over the last several years, we have been using advanced computational ecology methods and the tenets of open science to attempt to predict the probability of cyanobacteria blooms and provide access to the tools and data we develop for others to build upon. While it may be clear to us, who spend several hours a day behind a computer screen thinking about computer code and calculating uncertainty, that this work is important, most people do not know what computational ecology is. In addition, they don't understand how computational ecology can be used to establish an adequate understanding of the inherent ecosystem uncertainty that might help us better manage lakes to reduce cyanobacteria bloom risk. The purpose of this article is to introduce the concepts of computational ecology and open science and describe why we think they will advance our understanding of cyanobacteria blooms and help us make better predictions.

What is Computational Ecology?

Computational ecology is an interdisciplinary field that takes advantage of modern computation abilities to expand our ecological understanding. As computational ecologists, we combine data sets and advanced statistical/mathematical computational methods to build models that often cover broad spatial extents. This field is also fully entrenched in an ethos of open science and scientific reproducibility. Computational ecologists must have diverse skills as we are required to master data management and curation, coding, data analysis, and visualization, in addition to our ecological expertise. Essentially, we use big computers and big data to move ecological understanding forward.

The computational ecologist's toolbox works well for exploring the complexity of cyanobacteria-related questions. All areas of ecology are complex, but this complexity increases when dealing with the cyanobacteria phylum. This phylum has high species diversity and yet the individuals are small in physical size. If you want to study polar bears, it's fairly straightforward; you count polar bears. If you want to study cyanobacteria, what do you count? How do you count? It's not that these questions don't have answers, it's that most cyanobacteria experts answer the questions in different ways. This results in substantial data uncertainty. Each method used to measure cyanobacteria has its pros and cons. This, of course, means that the models from different data sources have to be interpreted according to the limitations of the cyanobacteria data. And we haven't even talked about the complexity involved in measuring environmental response variables and if those variables are ecologically meaningful to cyanobacteria.

Given the complexity and size of the data we must look outside traditional statistical methods to analyze our data. One of our favorite computational methods is "random forest," which we use frequently to build classifier models. This method is a machine learning approach that allows us to make robust predictions from large amounts of data with multiple data types. The random forest algorithm partitions the data into training and test data sets. Then the data are hierarchically partitioned into increasingly more homogenous groups based on a subset of the environmental variables. The test data set is then used to measure how well we did. This process is repeated multiple times to ensure that we have captured the true signal of the data.

For our most recent work, we used the U.S. EPA's National Lake Assessment (NLA) data from 2007 to build random forest models of lake trophic status. The NLA is a probabilistic sampling of 1,000+ lakes across all eco-regions in the continental US (Figure 1). Lake trophic status was used as a proxy for cyanobacteria abundance. We can do this because we know that cyanobacteria abundance and chlorophyll-*a* concentrations (which are used to classify lake trophic status) are positively correlated. In other words, the amount of cyanobacteria in a lake tends to increase as the amount of chlorophyll-*a* increases. There are several advantages to using lake trophic status as a proxy for cyanobacteria. By using chlorophyll-*a* concentration based trophic status, we are not constricted by one measure of cyanobacteria. While at the same time, we are using a unit that has real world meaning to lake managers. These broad trophic state classifications are good predictors of ecosystem health, which directly relates to ecosystem services/disservices (e.g., potential for recreation, good aesthetics, and fisheries).

The gold standard for understanding cyanobacteria in lakes is direct measurements of water quality variables, such as levels of nutrients, chlorophyll-*a*, and pigments. This requires the ability to take on site ("in situ") samples; something that cannot realistically be done for every lake in the country. Our modeling work

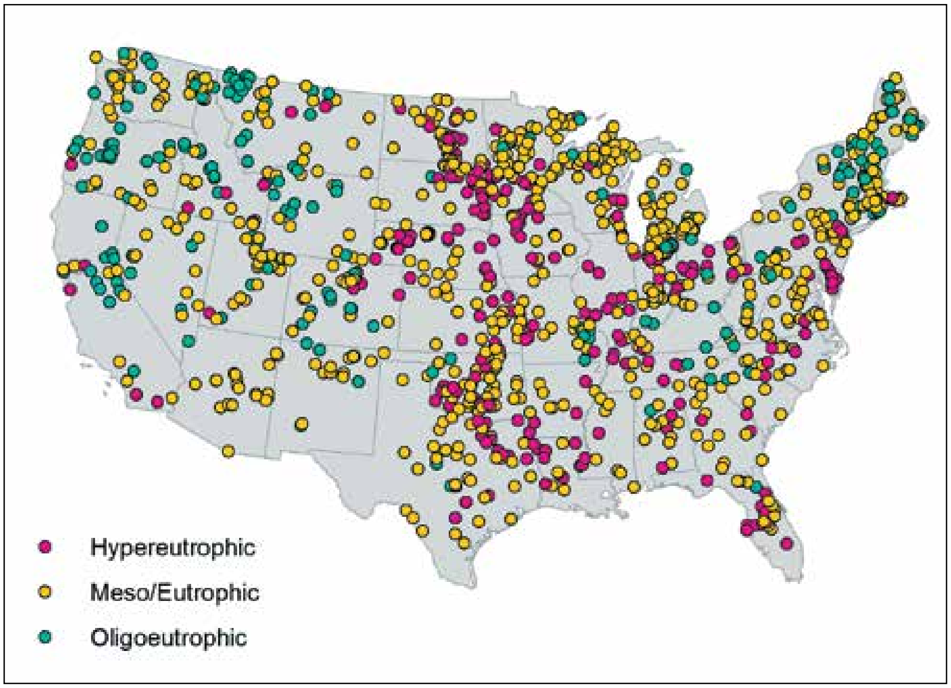


Figure 1: Map of the 2007 National Lake Assessment survey locations. Points are color-coded according to lake trophic status.

is focused on predicting cyanobacteria bloom risk for lakes that have not been directly sampled. Empirical data from lakes are combined with remote sensing and geographic information systems (GIS) data to model bloom risk; results from this work can then be extrapolated to all lakes in the continental United States. The work is starting to shed some light on landscape factors that may contribute to elevated bloom risk (Figure 2). For example, we know that different regions of the United States have different probabilities of bloom occurrences. We are also learning how lake morphometry, as well as the surrounding land use, impact lake trophic status.

As our work progresses, we are increasing the complexity of our modeling efforts by developing a Bayesian multilevel model. This approach offers numerous exciting advancements for cyanobacteria predictions. First, we are moving from using lake trophic status as a proxy for cyanobacteria to directly modeling microcystin, a common cyanobacteria hepatotoxin. We used the results of random forest modeling to select variables for inclusion in a Bayesian multilevel model of microcystin concentrations (Figure 3). Bayesian statistical methods start with prior beliefs and combines these with new information from the experiment, represented by the likelihood function, to form the posterior beliefs for the model parameters (Hoff 2009). Model parameters under a multilevel modeling framework are eco-region specific, but they are also assumed to be exchangeable across eco-regions for broad continental scaling (Gelman and Hill 2006; Qian et al. 2010). The exchangeability assumption ensures that both the common patterns and eco-region specific features will be reflected in the model. Furthermore, the method incorporates appropriate uncertainty estimates. This modeling approach has the added benefit of allowing us to update our assumptions when we have new data. And since the NLA is repeated every five years, we will be able to improve our base knowledge once the newest NLA data are released publically.

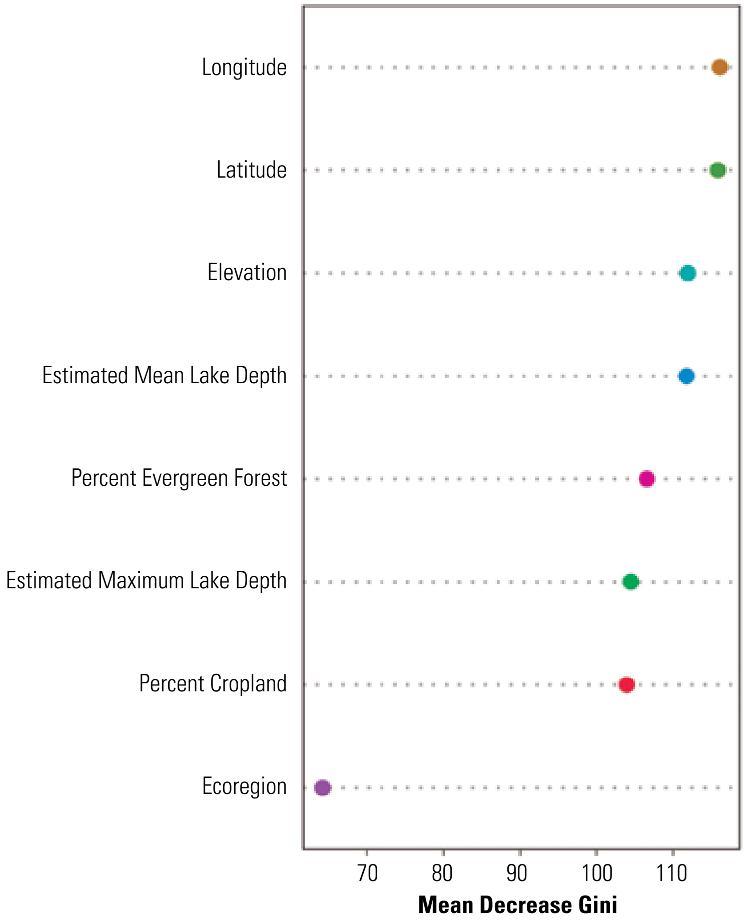


Figure 2: Plot of ranked mean decreased Gini from the random forest model predicting three levels of lake trophic status. This model's predictions were based solely on GIS-derived variables. Essentially, this figure illustrates the order of variable importance in development of the model. The most important variables are longitude and latitude, which would lead us to conclude that there is a spatial gradient across the U.S. We also can conclude that there is a gradient along elevations, because elevation is the third-most important variable.

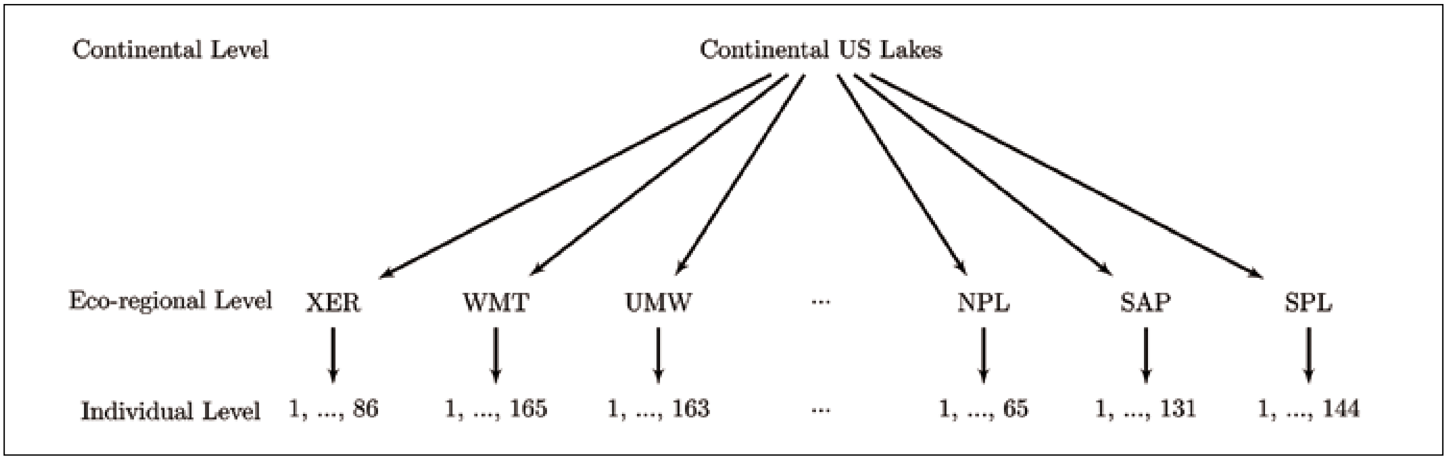


*Figure 3: Structure of the Multilevel Model: The U.S. continental, highest level, is divided into nine eco-regions. The eco-regions are divided into individual lakes (1148), lowest lev*el.

What is Open Science?

All of this modelling would be impossible without open access to modern computational methods and the data that support our models. Broadly speaking, this is often referred to as "open science." This broad area has been defined as having several components. These components suggest that "open science":

- is transparent (and, of course, open)
- includes all parts of research (data, code, etc.)
- allows others to repeat the work
- should be posted on an open and accessible website (while protecting personally identifiable information, etc.)
- occurs along a gradient (i.e., not just a binary open vs. not open)

At the EPA, we are learning how to make our research on cyanobacteria and human health meet these criteria. We are implementing open science in three ways: (1) making our work available via open access publishing; (2) providing access to the code used in our analysis; and (3) making our data publically available. The goal of these efforts is to increase the reproducibility of our work, reach broader audiences, and eventually have a greater impact on society's understanding and management of harmful algal blooms (HABs). Specifically, we are using the following open science channels to benefit the management of harmful algal blooms:

Open access publications: Our traditional venue for sharing research is peer-reviewed publications. A problem with many journals is that these articles are only accessible to those who have paid to gain access. An increasingly common option is to publish in open access journals or to pay additional fees to make sure a given article is open access. Researchers in our group consistently use open access venues for our research. By taking this step we are able to reach a much broader audience with our work.

Open source software: As computational ecologists we rely on scientific software to conduct our work. A very important part of using this software is to be able to check that the methods encoded in this software are valid. The only way to do this is to use software that is open source (i.e., the code is available to review, enhance, or modify). Not only do we use open source software such as the R Language for Statistical Computing (http://www.R-project.org/), we also contribute back to the open source community. Members of our group have developed software to support modelling of lakes (e.g., lakemorpho package [http://cran.r-project.org/web/packages/lakemorpho/index.html]) and we actively use the U.S. EPA's organizational account on Github (https://github.com/USEPA) for collaborating on code development and sharing other aspects of our work. By providing open access to our computational methods we allow others to repeat the same analyses or build from it.

Open data: The last area where we are just now starting to work is providing access to our data. As mentioned, we strive to publish our work as open access and along with those publications, we have, when possible, made the datasets that support that work available via supplemental materials. More recently we have released a first version of a national lake morphometry dataset. Those data are available, as GIS files, from https://edg.epa.gov/clipship/ under the heading, "National Lake Morphometry." We plan to continue improving this dataset.

Going Forward

We have been using computational ecology and open science in our HABs related research for several years now and have many plans going forward. First, we are expanding our modelling efforts to include new methods and endpoints. We hope to work with managers to identify what qualities of freshwater HABs are most important to predict. Second, new data are always becoming available that can inform our work. For instance, the 2012 National Lakes Assessment data (http://water.epa.gov/type/lakes/lakessurvey_index.cfm) should be available in the very near future, citizen science efforts such as those done by Rhode Island's Watershed Watch program have a rich trove of data that can help us better model HABs, and new cyanobacteria monitoring programs are starting to come on-line (see "New England Region Cyanobacteria Monitoring Program" in this issue). All of these will provide a fresh look at the HAB problem. In addition to these data sets, we are planning the development of a national lake database. Our lake morphometry data are the first step but we envision a database with multiple sources of water quality data, a mechanism for updating the data and access provided in a variety of ways for a variety of users. In short, understanding the dynamics of lake trophic status and cyanobacteria bloom risk is an increasing concern for lake resource managers. The computational approaches we describe here, as well as conducting research via the tenets of open science, will allow us to make significant advances in cyanobacteria ecology and other related fields.

Selected References

Breiman, L. 2001. Random forests. *Machine learning* 45, 5-32.

Hoff, P.D. 2009. A First Course in Bayesian Statistical Methods. Springer Science & Business Media. Springer.

Gelman, A. and J. Hill. 2006. Data Analysis Using Regression and Multilevel/hierarchical Models. Cambridge University Press, 2006.

Paerl, H.W. and T.G. Otten. 2013. Harmful cyanobacterial blooms: causes, consequences, and controls. *Microbial ecology,* 65(4): 995-1010.

Pascual, M. 2005. Computational ecology: from the complex to the simple and back. *PLoS computational biology,* 1(2): e18.

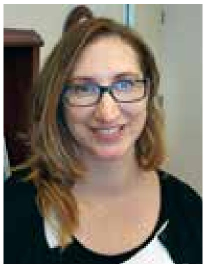
Qian, S.S., T.F. Cuffney, I. Alameddine, G. McMahon and K.H. Reckhow. 2010. On the application of multilevel modeling in environmental and ecological studies. *Ecology*, *91*(2): 355-361.

Betty Kreakie, Ph.D., is a research ecologist for the U.S. EPA's Office of Research and Development in Narragansett, RI. Her work focuses on the development of spatially explicit, landscape level models that predict how biological populations and communities will respond to anthropogenic influences such as nutrient and contaminant inputs, climate change, and habitat conversion. You can contact Betty at Kreakie.betty@epa.gov.

Jeffrey Hollister is a landscape ecologist with expertise in the spatial component of ecology and environmental sciences. Since May of 2006, he has worked as a research ecologist with the U.S. EPA's Atlantic Ecology Division in Narragansett, RI. His current research focus is on how nutrients drive risk of cyanobacterial blooms in lakes and ponds. A unifying theme to his research is using Open Science (Open Access, Open Source, and Open Data) to benefit environmental science.

Farnaz Nojavan is an ORISE Postdoctoral Fellow, at the U.S. EPA Atlantic Ecology Division. She is broadly interested in ecological modeling, aquatic ecosystems, Bayesian statistics, risk assessment, and environmental decision analysis. A central theme in her interdisciplinary research is the use of Bayesian statistics to improve inference and prediction. Her current research draws upon Bayesian multilevel modeling and datasets from disparate sources to investigate cyanobacteria distribution, microcystin risk, and changes in the algal community in lakes of the continental United States. You may reach Farnaz at: Nojavan.Farnaz@epa.gov.

Bryan Milstead is a research ecologist with the U.S. EPA in Narragansett RI. Bryan has worked extensively throughout South and North America and the Caribbean on varied research projects involving a dazzling variety of habitats and organisms. The one theme that holds all his work together is a strong interest in quantitative analysis and data management. He uses the open source R programming environment for most of his work in modeling, statistics, data manipulation, GIS, and graphics.



Lahne Mattas-Curry is a public affairs specialist in EPA's Office of Research and Development's Office of Science Communications. She brings nearly two decades of public relations and strategic communications experience to EPA. You can contact Lahne at mattas-curry.lahne@epa.gov.

(GILINSKY . . . Contintued from page 12)

Providing clean and safe water for healthy, thriving communities will require new solutions. Shifting rain patterns and seasonal temperatures across the country, in combination with increasing nutrient pollution, can lead to increases in harmful algal blooms. The science on harmful algal blooms is evolving and so are our solutions. Continued monitoring and treatment, and investment in our nation's water infrastructure, are necessary to prevent more blooms in the future**.**

I am encouraged by all of the great efforts going on at EPA and with our federal and state partners. When we all work together, we can adapt to new circumstances and protect our most precious resource for our children and our communities.

Ellen Gilinsky has served, since 2011, as the Senior Policy Advisor for Water at the Environmental Protection Agency. In this position Dr. Gilinsky addresses policy and technical issues related to all EPA water programs, with an emphasis on science, water quality, and state programs. Prior to this appointment she served as director of the Water Division at the Virginia Department of Environmental Quality (DEQ), where she supervised a diverse array of water quality and quantity programs, and before that as manager of the Office of Wetlands and Water Protection, helping to craft Virginia's non-tidal wetlands regulations and permitting program.