Cyanobacteria Dominance in Lakes and Evaluation of Its Predictors: a Study of Southern Appalachians Ecoregion, USA

Sajeela Ghaffar[1], R. Jan Stevenson[2] andZahiruddin Khan[3]

[1] *Institute of Environmental Sciences and Engineering, School of Civil and Environmental Engineering (IESE-SCEE), National University of Sciences and Technology (NUST), sector H-12 campus, Islamabad, Pakistan*

[2] *Center for Water Sciences & Department of Integrated Biology, 203 Natural Science Building Michigan State University, East Lansing, Michigan 48824*

[3] *Institute of Environmental Sciences and Engineering, School of Civil and Environmental Engineering (IESE-SCEE), National University of Sciences and Technology (NUST), sector H-12 campus, Islamabad, Pakistan*

** Corresponding author: sajeela_g@yahoo.com; Tel no. 092-323-5049722*

Abstract: Owing to their ecological and economic implications, the abundance of cyanobacteria has been an important subject in aquatic ecology. There has been a debate about which nutrients are a better predictor of cyanobacteria abundance, total nitrogen (TN), total phosphorus (TP) or nitrogen to phosphorus ratio (N:P). A classic study by Downing and coworkers in 2001 concluded that total N and total P are better predictors than TN:TP ratio. We picked lakes from a different region of the US, the Southern Appalachians (SAP), to test the same hypotheses using the National Lakes Assessment (NLA, 2007) database. This dataset consists of 116 selected sampling sites throughout the ecoregion. In this study, total cyanobacteria abundance was related to three predictors (N, P and N:P) using linear regression, non-parametric statistics and boosted regression trees (BRTs). Total N and total P were more strongly correlated to cyanobacteria abundance than N:P. Partial dependence plots from BRT analysis confirmed results of correlations showing higher relative influence values of TN and TP on cyanobacteria abundance. An important observation from both analyses was that cyanobacteria abundance increased rapidly with increasing TP above 5 µg TP/L. This is a similar response relationship between abundance of cyanobacteria and TP as reported by Downing et al. (2001). These TP thresholds for cyanobacteria dominance are important for TP management targets to control blooms in lakes.

1Introduction

The algal blooms in lakes, rivers and reservoirs are becoming a major problem worldwide [1]. There are many reasons as to why this is a problem. First is a decline in quality of water, especially freshwater bodies that serve as a source of drinking water for a large population. The second main reason is the ecological disturbance that hinders the normal balance between biotic and abiotic factors of an aquatic ecosystem. Algal blooms make the water body eutrophic which means there is an increase in the primary production which eventually spoils the structure and function of aquatic ecosystems, which has socio-economic implications [2], [3], [4] and [5]. The problem worsens when harmful toxin-producing algal blooms of cyanobacteria or cyano-HABs grow and produce toxins that threaten human and ecosystem health [6] and [7].

Although hydrological and climate factors play a part in proliferation of cyanoHABs [8] and [9], human activities such as urbanization, industrialization and rising agricultural activities due to ever increasing population are the main cause of high wastes. This waste when thrown untreated into water bodies, leads to high levels of nitrogen, phosphorus and other nutrients leading to higher algal growth [10]. More than 50% of phosphorus comes from human waste and 20%–30% from detergents [11]. Animal feedlots are sources of both nitrates and phosphates. As the high levels of nutrients (nitrogen and phosphorus mainly) enter water bodies, they lead to negative ecological costs to aquatic ecosystem; furthering growth of cyanobacteria; high nutrient turnover; loss of ecosystem functions and biodiversity; and decline of water quality and public health [12] and [13].

Since cyanoHABs have and are expected to decrease the water usability for potable and recreational purposes [14], their management is very important from ecological as well as economical point of view. Although various physical and chemical factors play a

part in cyanobacteria blooms, identifying the factor contributing the most or the limiting factor is key. This leads to decisive and smart management practices, rather than unsystematically trying to abate nutrients or any physical factor. One of the first steps for identifying the most limiting factor is the finding factor most highly related to cyanobacterial abundance. There has been a debate about the best predictor for cyanobacteria growth, and different studies narrowed the focus to TN, TP and TN:TP (for example [15], [16], [17] and [18]). The conclusions and suggestions vary. A classical 37 year experiment on Canadian lakes by [19] showed phosphorus directly controlling algal blooms. Another extensively cited study by [15] suggested ambient N:P influencing cyanobacteria growth. Many such experiments were later debated for their data collection approach, sample size as well as theoretical basis as more and more studies followed up the classical ones. Most studies have commented and discussed the role of each in theory from surveys and experiments. Few have used an objective survey of nutrient chemistry and cyanobacteria abundance in a large number of lakes as a line of evidence for the limiting factor.

In this study, we aim to identify the most important and efficient predictor for cyanobacteria blooms in freshwater bodies. The rationale being to develop management practices and assigning total maximum daily loads (TMDLs) for water bodies by identifying the nutrient concentration at which risk of high abundances cyanobacteria abundances increase. We aim to further our knowledge and understanding by testing linear regression, non-parametric statistics as well as a relatively new and robust modeling approach called boosted regression trees (BRTs). Boosted regression trees are particularly valuable for characterizing interactions among multiple variables, non-linear relationships and potential thresholds in cyanobacteria abundance along nutrient gradients. Thresholds in ecological responses are particularly valuable for developing stakeholder consensus for management targets [20].

2 Methodology

The data from National Lakes Assessment (NLA) study carried out by US Environmental Protection Agency (EPA) was used for this project [21]. NLA is a large and comprehensive dataset consisting of physical, chemical and biological studies of water bodies across the United States. The US is divided into nine ecoregions in the study. We selected the Southern Appalachians ecoregion for the project as a representation of warm temperate climate region. This subset of data includes 116 lakes.

We tested three predictors with potential influence on cyanobacteria growth, i.e. TN, TP and TN:TP. We chose "abundance" as a measure of cyanobacteria growth as response variable. The selection of variables was based on past studies where these were deemed to be the most important or influential ones for cyanobacteria prediction. Microsoft Access (2013) was used as the tool for extraction of parameters from various datasheets and appending them into a single excel sheet.

Linear models were tested for cyanobacteria abundance as a function of selected predictors. All analyses including the subsequent ones were executed in R ver. 3.0.3 [22]. Residual standard error (RSE) was used as a measure of accuracy i.e. the closeness of fit to the points (for example its value is 0 when model fits to the data perfectly, though this would be most probably due to overfitting). The linear regression analyses were then compared to BRT results.

General trends in relationships of predictors and cyanobacteria abundance were characterized by LOWESS smoothing. LOWESS is a non-parametric regression approach producing a scatter diagram that shows the relationship between predictor and response variables over localized subsets of data, hence the term "local" as opposed to "global" fitting. This is especially useful to visualize patterns in the data as there are no assumptions defining the relationship. The plots were constructed using a tension factor (α) of 0.3. The response variable, cyanobacteria abundance, was log transformed with base 10. The predictor variables were also log transformed to get a better visualization of plots as they expand over a large range and tend to cluster on left hand side of the plot if untransformed.

BRTs are a combination of boosting and regression trees. Boosting is a machine learning (ML) approach, based on the concept that many weak learners make a powerful learner. Regression trees belong to classification and regression tree (CART) assembly of models [23]. This is a relatively novel approach and with strong ML basis that is finding more and more uses in ecological studies due to its high predictive power, better precision and accuracy compared to other modeling approaches [24]. BRTs work by minimizing predictive deviance obtained by fitting successive regression trees. The trees are built by an iterative learning process of modeling interactions between predictors and random selection of observations in each iteration [25] and [23]. The results of BRT models are viewed by partial dependence plots (PDPs) that show relationship as well as relative influence of each predictor on the response variable. The algorithm estimates relative influence by taking into account the number of times the specific variable is used to split tree nodes, given weight by squared improvement at each step, and then averaging it over all trees [26].

BRT analysis was run using the script from [23], a package called gbm – gradient boosting machines. The package also includes 10-fold cross validation in which the data is divided into 10 parts. 9 parts are used for training and 1 part is used for testing. This is an effective way of validating the model by performing analyses on multiple iterations of the dataset.

The analysis was run with setting a bag fraction of 0.5 which means 50% of training data is randomly picked for each iteration; and a tree complexity of 3 that means up to 3 way interactions of predictors can be modelled. A learning rate of 0.001 was set for all analyses, which is a parameter determining contribution

from each tree to build the next in making final decision. All variables including the response and predictors were log transformed as convergence could not be achieved over the large range of non-transformed values. 10 fold cross validation was used to validate the model, which is a commonly used approach. PDPs were constructed from the same gbm package, which includes script for the purpose. Predictor selection was then refined by observing PDPs based on relative influence of each predictor.

Model comparison was done by cross-validation (CV) correlation values. The BRT model with combination of predictors giving minimum predictive deviance and maximum explained error variance values was selected. The model was tested for performance on a subset of data withdrawn earlier. The same model parameters were used and cyanobacteria abundance was predicted on new data. The model performance was verified using root mean square error (RMSE) values between observed and estimated values, which is basically the square root of variance of residuals, having same units of response variable and representing accuracy of model for predicting values.

3. Results

The linear regression analysis (summarized Table 1) indicated TN:TP as the weakest predictor of cyanobacteria abundance among selected three (R2 = 0.14, RSE = 0.88). TP came out to be the best predictor with highest R2 (0.28) as well as lowest RSE estimates (0.803) indicating better prediction accuracy. TN also has higher R2 (0.184) and lower RSE (0.857) than TN:TP indicating a relatively stronger relationship. All analyses were significant at $p<0.001$. In summary, TP is related linearly stronger to cyanobacteria abundance than any other predictor, especially TN and TN:TP.

Table 1. Linear regression statistics showing ability of three selected predictors to predict cyanobacteria abundance in Southern Appalachians ecoregion. (RSE on df = 128; all analyses are significant at <0.001)

Variable	Linear model equation	RSE (df 128)	R^2
TP (µg/L)	Abundance = 2.687 x TP + 0.896	0.803	0.28
TN (µg/L)	Abundance = 1.21 x TN + 0.659	0.857	0.184
TN:TP	Abundance = -0.861 x TN:TP + 4.94	0.88	0.14

LOWESS plots (Fig. 1) show little nonlinearity in relationship between cyanobacteria abundance and the selected predictors. The LOWESS plot for TP (Figure 1a) shows that cyanobacteria are abundant even at lowest observed TP concentration (1 µg/L); at around 5 µg/L TP, the abundance of cyanobacteria starts increasing.

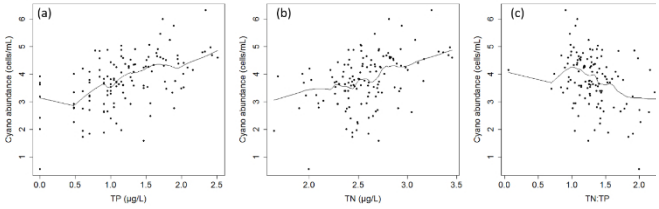

Figure 1: Cyanobacteria abundance in the lakes of Southern Appalachians ecoregion as a function of three selected predictors (a) Total Phosphorus (µg/L) (b) Total Nitrogen (µg/L) (c) Total N:Total P. Scatter represents the observations; solid lines are LOWESS trend lines (α=0.3). All values log-transformed to even out the spread.

The BRT model relating cyanobacteria abundances to TN, TP, and TN:TP was achieved after fitting 1500 trees and explained about 63% and 45% (SE 0.051) of variance in the entire training data set and with cross validation respectively. The partial dependence plots (PDPs) showed that TP had a highest relative influence (41.9%) on cyanobacteria abundance. TN followed TP with a relative influence value of 35.4%. TN:TP had 22.6% influence on cyanobacteria abundance (Figure. 2).

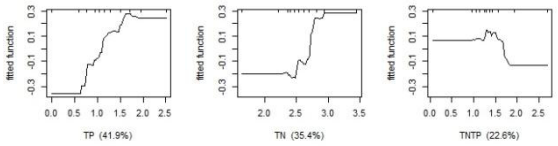

Figure 2: Partial Dependence Plots (PDPs) obtained from boosted regression tree model using 3 predictor variables

RMSE value for training data was 0.73 and for test data was 0.77. If test data RMSE for test data is much larger than training data, it suggests that model validation is poor and it does not have much predictive importance. Since these values came to be quite close, it is indicative of a good predictive performance. The BRT plots show strong nonlinearity between cyanobacteria abundance and nutrient concentrations or ratio. The TP plot shows a sudden rise in abundance at around 5 µg TP/L and saturation around 60 µg TP/L. TN plot shows a relatively sharp rise in cyanobacteria at about 130 µg TN/L. BRT results for TN:TP show that cyanobacteria abundance start to decline at a TN:TP of about 15.

4. Discussion

We aimed to understand the response of cyanobacteria to selected predictors with emphasis on choosing the best one for cyanobacteria prediction from a waterbody management view point. The data included observations over a broad phosphorus and nitrogen range. We expected to see relationships between

cyanobacteria abundance and selected predictors and identify the strongest related ones and the most effective nutrient range for cyanobacteria growth. We also expected to see cyanobacteria responses to nutrients similar to those reported by [17]. The results from Southern Appalachians show some interesting trends with a few in-line with [17] findings and a few giving a more thorough insight of Southern Appalachians nutrient dynamics.

Linear models for cyanobacteria and either TN, TP or TN:TP ratio in combination with corresponding LOWESS plots give a few more insights. While all these variables relate to abundance, TP displays the strongest relationship. TN follows TP with an R2 about 35% smaller than that of TP. As for the nitrogen to phosphorus ratio, cyanobacteria are more abundant in lower range of TN:TP; as it goes beyond 15:1, cyanobacteria become less and less abundant. The decline stops around 60:1 and then stays almost constant. This follows resource ratio competition theory that suggests that cyanobacteria have a tendency to thrive well in nitrogen limited conditions [27] and where N:P ratios are lower than 20:1 [28] and [29]. This is attributed to the fact that many cyanobacterial species are capable of utilizing molecular nitrogen [30]. TN:TP however, is the poorest predictor if seen compared to nutrient concentrations (TN and TP) in our results. It shows the lowest R2 value amongst all three which is 50% lower than that of TP and 22% than TN. In addition, TN:TP shows highest RSE; about 8% more than TP and 3% than TN.

The BRT model also shows TP having the higher relative influence on cyanobacteria abundance than both TN and TN:TP. A few points are noteworthy here. First is a long discussed concept about whether the nutrient concentrations (TN and TP) or the ratio TN:TP is a more efficient predictor of cyanobacterial blooms (for example [31], [17], [32], [33],[34] and [35]). Results from this study favor the latter notion based on linear regression, LOWESS and BRT models supported by PDPs. TN and TP clearly outperform TN:TP in predicting cyanobacterial abundance.

An important point from the results is the pattern of response that cyanobacteria abundance shows to total phosphorus concentration. The BRT partial dependence plot (Fig. 2) show that around a TP concentration of 5µg/L, cyanobacteria abundance is high until around 100µg/L TP where almost 90% of maximum cyanobacteria abundance is observed. It is seen to decline afterwards. This range is in line with that reported by [17]. It is suggestive of capability of cyanobacteria to grow well at low TP concentrations. This is most likely owing to several adaptive features of different cyanobacterial species that include nitrogen fixation and capability to store phosphorus in the cells that can be utilized for growth as per need. The practical implication of this observation is related directly to management of water body. It implies that accounting for other predictors, ~5 µg/L TP is a crucial point for increase in probability of cyanobacteria blooms and unless TP is reduce below 60 µg TP/L, little reduction in cyanobacteria abundance should be expected.

It may seem plausible looking at P dependence of cyanobacteria, that adjusting P only can be effective for cyanobacteria control in waterbodies. However, in a watershed where nitrogen influx is high, controlling P alone cannot control non-nitrogen fixing cyanobacteria in the water body. This is especially important considering the dramatic increase in nitrogen influx in water bodies [36]. Reference [33] suggest that in lakes where non-nitrogen fixers are abundant, regulating only P might not be very effective. For example, Microcystis can utilize phosphorus very efficiently by vertical migration, stock phosphorus from sediment-water interface and can reemerge to bloom. This is especially true for shallow lakes where sediment-water interface is easy to get to along with sufficient light penetration and water mixing. Reference [37] also advise that effective control can only ensue by rigorous nutrient management owing to nonlinear nature of association amongst environmental factors and cyanobacteria growth [36]. This study points out that TN and TP are the important predictors that should be given more importance than TN:TP. Although the relationship of the latter is substantiated by resource ratio theory, TP and TN deliver a simpler understanding and a more efficient approach to control cyanobacteria blooms. This is also supported by the fact that not all cyanobacteria have the ability to fix nitrogen nor are they a great competitor for nitrogen [17].

Acknowledgment

The funding to conduct research work at Michigan State University (MSU), East Lansing, Michigan USA was kindly provided by National University of Sciences and Technology (NUST), Islamabad, Pakistan; this study is a part of that research. The authors also acknowledge USEPA whose comprehensive NLA data is openly available and was used for this study.

References

1. X. Yang,X. Wu, H. Hao, Z. He, Journal of Zhejiang University Science B.**9**(3):197-209 (2008)
2. B. Mortazav, R.L. Iverson, W.M. Landing, G.G. Lewis, W. Huang, Marine Ecology Progress Series, **198**:19– 31 (2000)
3. E.J. Phlips, Algae and eutrophication. In Bitton, G. (ed.), Encyclopedia of Environmental Microbiology. Wiley, New York. (2002)
4. E.L. Bledsoe, E.J. Phlips, C.E. Jett, K.A. Donnelly,Ophelia **58**:29–47 (2004)
5. I.T. Webster, N. Rea, A.V. Padovn, P. Dostine, S.A. Townsend, S. Cook, Marine and Freshwater Research **56**:303–316 (2005)
6. J.H. Landsberg, Reviews in Fisheries Science 10: 113– 390 (2002)
7. W.W. Carmichael, A world overview one-hundred, twenty-seven years of research on toxic cyanobacteria—Where do we go from here? In: Hudnell, H.K. (ed.) Cyanobacterial Harmful

Algal Blooms: State of the Science and Research Needs. Advances in Experimental Medicine & Biology, Vol. 619. Springer. 500 pp (2008)
8. Y.S. Li, X. Chen, O.W.H. Wai, B. King, Water Environmental Research,**76**: 2643–2654 (2004)
9. E. Jeppesen, M. Sondergraard, M. Meerhoff, T.L. Lauridsen, J.P. Jensen, Hydrobiologia**584**: 239–252 (2007)
10. M. Schrope, Nature, **452**(7183): 24–26 (2008)
11. F.A. Khan, A.A. Ansari, The Botanical Review, **71**(4):449-482 (2005)
12. W. Liu, R.L.Qiu, Journal of Chemical Technology and Biotechnology, **82**(9):781-786 (2007)
13. M.W. Suplee, V. Watson, M. Teply, H. McKee, Journal of the American Water Resources Association, **45**(1), 123-140 (2009)
14. S. Lyck, Journal of Plankton Research, **26**(7), 727-736 (2004)
15. V.H. Smith, Science, **221**: 669–671 (1983)
16. A.M. Trimbee, E.E. Prepas, Canadian Journal of Fisheries and Aquatic Sciences, **44**(7), 1337-1342 (1987)
17. J.A. Downing, S.B. Watson, E. McCauley, Canadian journal of fisheries and aquatic sciences, **58**(10), 1905-1908 (2001)
18. R. Quirós, The nitrogen to phosphorus ratio for lakes: A cause or a consequence of aquatic biology. Water in Ibero-America: From Limnology to Management in South America. CYTED XVII (Fernandez Cirelli A, ChalarMarquisa G, eds.). Madrid, Spain, 11-26 (2002)
19. Schindler, W. David, R.E. Hecky, D.L. Findlay, M.P. Stainton, B.R. Parker, M.J. Paterson, K.G. Beaty, M. Lyng, S.E.M. Kasian, Proceedings of National Academy of Sciences USA **105**:11254–11258 (2008)
20. R. Muradian, Ecological economics **38**(1), 7-24 (2001)
21. United States. Environmental Protection Agency. Office of Water. & United States. Environmental Protection Agency. Office of Research and Development. National lakes assessment a collaborative survey of the nation's lakes: draft. Washington, DC: U.S. Environmental Protection Agency, Office of Water and Office of Research and Development (2009)
22. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/ (2013)
23. J. Elith, J.R. Leathwick, T. Hastie, Journal of Animal Ecology, 77(4), 802-813 (2008)
24. G. De'ath, Ecology 88:243-251 (2007)
25. J.H. Friedman, Annals of statistics, 1189-1232 (2001)
26. J.H. Friedman, J.J. Meulman, Statistics in medicine, 22(9), 1365-1381 (2003)
27. I. Tõnno, The impact of nitrogen and phosphorus concentration and N/P ratio on cyanobacterial dominance and N2 fixation in some Estonian lakes. Tartu University Press (2004)
28. N.G. Bulgakov, A.P. Levich, Archivfürhydrobiologie, 146(1), 3-22 (1999)
29. K.E. Havens, R.T. James, T.L. East, V.H. Smith, Environmental Pollution, 122(3), 379-390 (2003)
30. H. Bothe, O. Schmitz, M.G. Yates, W.E. Newton, Microbiology and molecular biology reviews, 74(4), 529-551 (2010)
31. R.C. Lathrop, S.R. Carpenter, C.A. Stow, P.A. Soranno, J.C. Panuska, Canadian Journal of Fisheries and Aquatic Sciences, 55(5), 1169-1178 (1998)
32. R.W. Howarth, R. Marino, Limnology and Oceanography, 51(1part2), 364-376 (2006)
33. D.J. Conley, H.W. Paerl, R.W. Howarth, D.F. Boesch, S.P. Seitzinger, K.E. Havens,... &G.E. Likens,Science, 323(5917), 1014-1015 (2009)
34. W.M. Lewis Jr, W.A. Wurtsbaugh, H.W. Paerl, Environmental Science and Technology 45(24), 10300-10305 (2011)
35. D.W. Schindler, Proceedings of the Royal Society of London B: Biological Sciences 279(1746), 4322-4333 (2012)
36. H.W. Paerl, H. Xu, M.J. McCarthy, G. Zhu, B. Qin, Y. Li, W.S. Gardner, Water Research 45(5), 1973-1983 (2011)
37. H.W. Paerl, V.J. Paul, Water research, **46**(5), 1349-1363 (2012)

[a]Corresponding author: sajeela_g@yahoo.com