View Letter

Close

Date: May 26, 2015
To: "Jeffrey W. Hollister" hollister.jeff@epa.gov
cc: kmaloney@usgs.gov
From: "Freshwater Science" psb3@psu.edu
Subject: Freshwater Science Submission MS# 2015063

Ref.: Ms. No. 2015063
Modeling Lake Trophic State: A Random Forest Approach
Freshwater Science

Dear Jeffrey Hollister,

We have received and considered peer reviews of your submission. On the basis of the attached comments by the reviewer(s), as well as my own evaluation, I cannot recommend your manuscript for publication in Freshwater Science. I realize that this news is hard to hear, and I am sorry that the outcome was not better. In my Comments and Advice for Authors (see below), I have provided specific reasons for my decision and comments/suggestions that I hope will be helpful should you decide to submit your manuscript elsewhere. The referee comments (see below) also provide helpful suggestions.

Thank you for submitting your work to Freshwater Science. We will look forward to working with you again in the future.

Yours sincerely

Kelly Maloney
Associate Editor

On behalf of the Editors
Freshwater Science

Associate Editor's comments:

Similar to Reviewer #2 I found your use of the NLA data set to be of interest. I also found your use of the varSelRF package to be of potential use to future studies with similar data issues. However, both reviewers raised substantial concerns about the study, all of which I share. First, both reviewers questioned your use of discrete trophic categories and loss of information by doing so. Using the continuous data also could reduce your model set to just two: all variables vs GIS only. This would allow you to dive more deeply into the model results and patterns. Then, if you wish, you can classify into categories (as suggested by both Reviewers). Reviewer #2 provided an excellent suggestion to spatially display the residuals from the models. You should consider this comment, and perhaps a spatial map where the All and GIS only models differ, because they may help shed light on patterns. Both reviewers also suggested use of partial dependence plots to help visualize relationships. This is another good suggestion as the reader often wonders the functional form of the relationship; it would also lend support to your discussion. Both reviewers also raised concerns on supporting literature in the discussion, which I agree could be better supported with current literature. Both reviewers also question how and why the cyanobacteria analysis was included. This is a good comment because as currently written it appears to be an afterthought rather than importance piece. I had two additional comments. First, some of your models used ordinal data, but it was not clear if you used weighted model performance measures or not. Second, there has been debate on potential bias with variable selection in random forests (see Strobl et al. 2007); this may affect your inferences. Both reviewers provide some additional excellent suggestions that will hopefully help if you decide to submit elsewhere.

Strobl, C., A. L. Boulesteix, A. Zeileis, and T. Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. Bmc Bioinformatics 8.

Reviewers' comments:

Reviewer #1: Review: Modelling lake trophic state

This manuscript describes the application of random forest models to predict trophic state in US lakes using predictor variables that include in-situ measurements and landscape variables, and predictor variables that only include landscape variables. I had a few overarching comments, described below, followed by specific comments referenced to line numbers.

Overall, I found the statement of the goals of the manuscript to be too broad. In particular, it would be useful to frame the manuscript around a clearly stated research question or hypothesis. For example, can GIS variables provide as accurate a prediction of lake trophic status as in-situ variables? A goal stated as "Updating modelling efforts" is vague, and seems to suggest that the performance of the new method (i.e., RF) would be explicitly compared to the performance of more commonly used methods (i.e., linear regression). The research question ideally should also be a question that has not already been answered. Hence, the goal to describe variables that are important for describing lake trophic status does not seem appropriate as these variables have already been identified (e.g., nutrient concentrations, Secchi depth, chl-a concentrations, oxygen demand). Similarly, associations between trophic status and cyanobacteria biomass have been described, but the manuscript does not cite this literature. Finally, providing full access to data and analysis is nice, but is it really a research goal?

On a related note, the current results should be provided in the context of current literature, which requires a much more in-depth coverage of the literature than is currently represented in the Discussion. Many statements and assertions in the Discussion require one or more supporting citations.

In general, it seems to me that the analysis indicates that GIS variables do not provide predictions of trophic state that are as accurate as in-situ variables, and so the associated question that the authors should address is whether this loss in accuracy is acceptable given that predictions can be obtained for more lakes. How do the authors envision this model being used and not being used? For example, a model based only on GIS variables does not help with setting appropriate nutrient concentrations in lakes, but it may help target monitoring toward lakes that are likely to be over-enriched. Would it be possible to do case study in a region to show how the model might indicate lakes that require monitoring? Such an exercise would also help assess the final utility of the model. For example, if the best predictor variable is ecoregion, then the model essentially tells the manager that all lakes in a large ecoregion need more monitoring, an outcome which is not useful.

I also wondered about the choice of discrete trophic categories for the model response variables. Most practitioners model a continuous chl-a concentration (Carlson 1977) and then define categories based on the model results (e.g., discrete concentration ranges for TP, Secchi, and chl-a). The present approach of using pre-defined trophic categories seems to unnecessarily compress the information provided in chl-a measurements to 2-4 discrete categories. In addition to "losing" information in measured chl-a, the interpretation of the results depends somewhat on the choice of the boundaries between different trophic states, which are the subject of some debate.

Line 49: Add a citation for natural eutrophication.
Line 51: How does monitoring trophic state allow for identification of rapid shifts? Are you assuming that monitoring is frequent? If GIS variables only are used to identify trophic state, wouldn't they be mostly invariant with time, and therefore, not at all useful for observing rapid shifts?
Line 52: "unusually high" needs qualification. There are naturally hypereutrophic systems.
Line 54: What is "fouling"
Line 56-57: Consider using a different term than "first cut". Too colloquial.
Line 59: "efficient"
Line 66: Throughout the manuscript: avoid using "this" without a following subject noun.
Line 99: "utilized". Methods should be in past tense.
Line 111: Some description of the field methods is probably necessary. At the very least, a citation of a description of the field methods should be provided. USEPA (2009) does not include field methods.
Lines 124 -125: Please clarify which of these metrics are derived from model predictions and which are derived from field measurements. How was Fetch computed?
Line 128: "Random forests"
Line 137: "downfall" seems harsh
Line 143 -145: Awkward sentence.
Line 155 - 156: Consideration of 3 combinations of trophic state seems like a good idea, but I didn't see any discussion of the effects of the different combinations. Wouldn't we, by definition, expect poorer predictive performance as we increase the number of discrete classes? Is there some baseline change in predictive performance that we can measure against?
Line 194-197: This sentence and the thought it represents seems misplaced.
Line 209: How does Table 1 support the preceding statement?
Line 221 and 228: Why are the sample sizes for Model 1 and 2 different? I thought the only difference in the two

models would be in how chl-a was discretized.
Line 243: Why is sample size different from Model 1? Isn't Model 4 identical to Model 1 except that in-situ variables are dropped?
Line 276 - 291: This discussion of maximum probabilities is interesting, but demands its own methods description and a description of the results. Cramming everything into the Discussion makes it too difficult to follow.
Line 294, 297: "cyanobacteria abundance": Why the switch from trophic state to cyano abundance here?
Line 300 - 307: Discussion here is interesting but entirely unsupported by citations and/or results. A partial dependence plot of turbidity would at least show the shape of its relationship with tropic state.
Line 310: Wasn't ecoregion the only variable that wasn't selected in Model 4? Again, citations to relevant literature would be useful. It seems to me that the selected variables across the different models were quite different when predictors were limited to GIS variables, which may be an indication of the weakness of the relationships between the GIS predictors and lake trophic status.
Line 326: This section on cyanobacteria requires separate descriptions in the methods and results sections. As it stands it seems like something that has been added haphazardly to the end of the manuscript. As with my comment above, it seems that continuous models linking cyanobacteria abundance and chl-a would be far more interesting and informative to dig into further (Fig 12).
Line 356-357: Don't these results suggest that the models will provide accurate predictions for certain lakes and for other lakes the predictions may not be accurate?

Reviewer #2: I was excited to see the National Lakes Assessment data being used for more than just national summaries of lake condition. There have been a number of recent papers on modeling various lake water quality parameters using landscape and in situ predictors, but this is the first one to use nationally representative dataset. Also, the random forest model approach seems promising for this application. However, there are several problems with the paper that make it not ready for publication in my view. I encourage the authors to consider my critiques and suggestions and resubmit it to FS or another appropriate journal. An improved version of this paper would make an important contribution to understanding of drivers of lake water quality.

1. Why not model trophic state as a continuous variable? You could then present the results as predicted vs. observed, and include standard errors of predictions. You can always convert continuous predictions back to classes. Related, the four class trophic state models are already pretty coarse (i.e., compared with modeling trophic state as a continuous variable), so I don't see the value in presenting the three and two class models. It seems like they were included just to increase the description of model accuracy in the abstract.
2. It would be helpful to include univariate partial dependence plots for the most important variables for at least the four class models. I would also suggest evaluating bivariate partial dependence plots for selected interactions that you expect might be important (e.g., TP and TN). These plots would be more informative if you switch to modeling trophic state as a continuous variable. The plots would also provide the basis for a more nuanced discussion of the effects of predictor variables on trophic state, such as whether the relationships are linear or non-linear.
3. While it does not appear to be as important as P or N, K was selected in most of the "all variables" models. This is an interesting finding, and worth exploring more.
4. I suggest using figure 1 to map trophic state and then add a companion figure that shows residuals from the model. This would help you identify geographic areas where the model appears to be under- or over-predicting trophic state, and could help you identify unmeasured variables that might be worth pursuing in future modeling efforts.
5. The "Variable Selection and Importance" section of the discussion needs to be significantly expanded. This should be the meat of the paper, where the model structure is unpacked and compared to mechanistic expectations and previous studies. The section currently contains one reference, which is woefully inadequate.
6. Please provide a brief summary of the water quality sampling methods, number of samples, and timing.
7. The use of the 3 km buffer for summarizing land cover is not defended. Why not use watershed land cover?
8. The model descriptions in the methods (L 161-173) and results (L 220-265) would be better presented in a table.
9. The cyanobacteria analysis seems tacked on and doesn't add much. I would omit it from this paper and consider dealing with it in a more sophisticated way in a separate paper.
10. Need additional references to recent work on this topic, e.g., Filstrup et al., Limnology and Oceanography, 59(5), 2014, 1691-1703, and Cross and Jacobson, Lake and Reservoir Management, 2013, 29:1-12, and references in these papers.
11. The writing is disorganized. There are several places where sentences that should be grouped are separated and many descriptions of methods or results that are not clear. I encourage the authors to engage the help of a copy editor before resubmission.
12. Please include more details on the arguments used in the randomForest model, including ntree, mtry, replace, etc.
13. L 198. The description of the mean decrease in Gini is incorrect. It is based on node impurity, but is not a permutation test. Review Breiman and Cutler et al. for details.
14. Need better descriptions of variables in Appendix 1, in particular, which ecoregion classification was used?
15. The content and purpose of figure 8 are unclear.

****

Close