Random Forests and a Potential Function for the Chesapeake Bay

Christopher Holder and Anand Gnanadesikan
Department of Earth and Planetary Sciences, Johns Hopkins University

Introduction

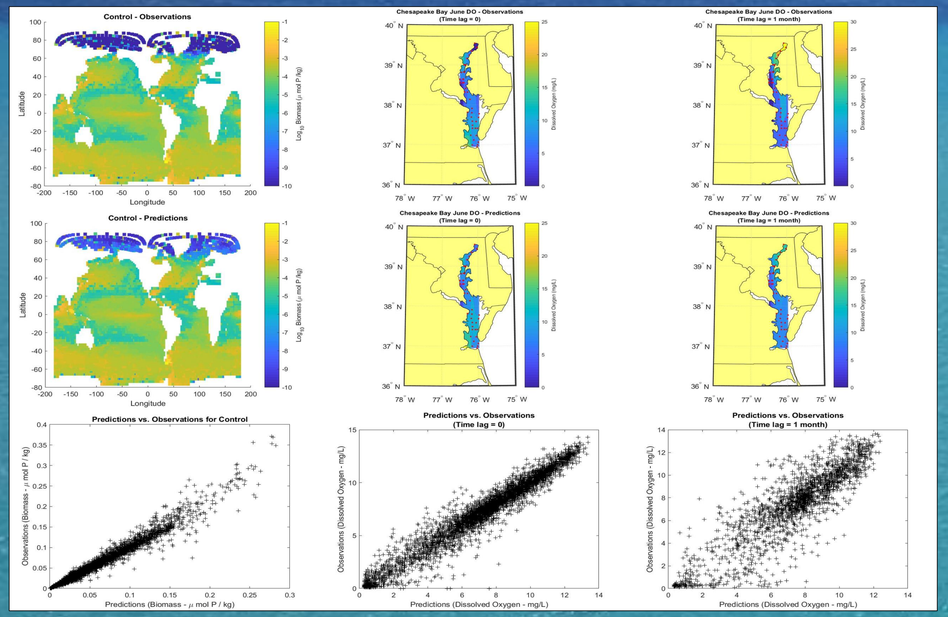
Low oxygen (anoxic) regions play a significant role in the health of the Chesapeake Bay ecosystem. Nutrient runoff from agricultural and industrial practices produce these low oxygen areas in the Bay, especially at depth during summer months. However, certain chemical components may have varying degrees of influence on dissolved oxygen (DO) in different regions of the Bay. Determining which combinations of nutrients and environmental variables produce these patterns can be difficult.

Machine learning has been used increasingly in science and private industry, but has yet to take hold in oceanography. A frequent criticism raised against these techniques is their "black-box" nature, in that it can be challenging to ascertain how outcomes were produced. Here, we examined a machine learning technique called random forest (RF) analysis for its ability to accurately predict outcomes on a biogeochemistry model and its interpretability. We then began a preliminary investigation of 41 Chesapeake Bay monitoring stations to assess if RF analysis could predict dissolved oxygen (DO) with no time lag and a time lag of one month.

Control Case

Our control case used model output for surface observations from a non-linear biogeochemistry model called BLING (Biogeochemistry with Light, Iron, Nutrients, and Gases) (Galbraith et al. 2010), in which the output is biomass and the inputs are iron, light, and nutrients. This was chosen as an ideal control scenario since the relationships are complex and they are known.

The RF analysis performed well, giving a R^2 of 0.972 between the predictions and observations. It was able to explain 96.5% of the variability and correctly predicted the most significant predictors as iron and light (irradiance). Additionally, the relationships via the partial dependence plots appeared to show that co-limitations were also captured in the analysis.

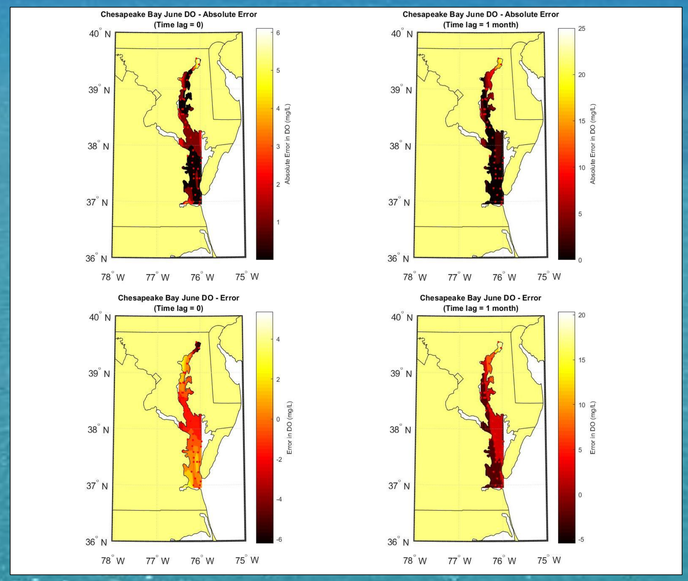


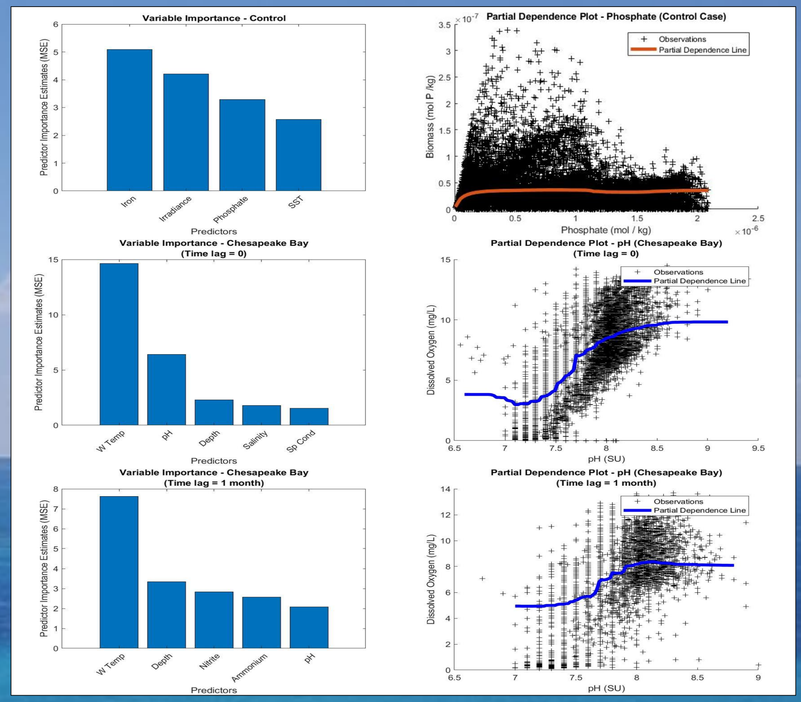
Chesapeake Bay

Data for 41 stations from the Chesapeake Bay was used in our preliminary RF analysis to predict dissolved oxygen concentrations with no time lag and a time lag of one month. The dataset included 28 variables from 2008 to 2018.

The initial RF model with **no time lag** used all possible predictors and produced a R^2 of 0.959 between the observations and predictions. Following this, we ran another RF model using only the top five predictors from the initial model. The new model was still able to give a R^2 of 0.943 and showed the top predictors for the "no time lag" model were water temperature and pH. The partial dependence demonstrated that water temperature had a negative relationship with DO, while pH had a positive relationship (shown). Furthermore, the only area whose pattern was not captured well was the Northernmost region near the mouth of the Susquehanna River where the observations were higher than the predictions.

The initial RF model with a **time lag of one month** again used all of the possible predictors to start. We used observations for the predictors from one month and used the subsequent month's observation for DO. The RF model gave a R^2 of 0.811 between the predictions and observations. Next, we ran a RF model with the top five predictors from the "one month time lag" model. This only reduced the R^2 to 0.771. The top predictors for this case were water temperature and depth. The "pH" variable was also seen again as one of the top predictors. The partial dependence plots showed a positive relationship between DO and pH (shown), while water temperature and depth showed negative relationships with DO. Similar to the "no time lag" model, this "one month time lag" model, showed that the area with the biggest differences between observations and predictions was near the mouth of the Susquehanna River.





Conclusions and Future Work

RF analysis does a remarkable job at capturing patterns within complex datasets, as shown by the results for the non-linear BLING control case. Furthermore, it appears to have some capability of predicting environmental conditions in natural datasets.

Future work may focus on a more in-depth analysis of these initial Chesapeake Bay results, as well as examining further time lags. Additionally, we may examine possible relationships between environmental conditions and microbial community genetics using RF analysis.

References

Breiman, L. Random forests, 2001. Machine Learning, 45(1):5-32.
Chesapeake Bay Program, Water Quality, Maryland Department of Natural Resources (2008-2018). *Chesapeake Bay Stations CB1.0 to CB5.3* [Data File]. Retrieved from http://data.chesapeakebay.net/WaterQuality.
Chesapeake Bay Program, Water Quality, Old Dominion University (2008-2018). *Chesapeake Bay Stations CB5.4 to CB8.1E* [Data File]. Retrieved from http://data.chesapeakebay.net/WaterQuality.
Galbraith, E.D., A. Gnanadesikan, J.P. Dunne, and M.R. Hiscock, 2010. Regional impacts of iron-light colimitation in a global biogeochemical model. Biogeosciences, 7(3):1043-1064.

Acknowledgements

The authors would like to thank the Johns Hopkins University, Department of Earth and Planetary Sciences and the National Science Foundation Integrative Graduate Education and Research Traineeship.