# Updates on Using Multi-media Modeling to Investigate Conditions Leading to Harmful Algal Blooms

Christina Feng Chang[1], Catherine Nowakowski[2], Marina Astitha[1], Valerie Garcia[3], Penny Vlahos[4], Chunling Tang[3], David Wanik[1], Jun Yan[5]

[1]Civil and Environmental Engineering, University of Connecticut, Storrs, CT, USA
[2]University of Rhode Island, RI 02881, USA
[3]National Exposure Research Laboratory, Office of Research and Development, US Environmental Protection Agency, Research Triangle Park, NC 27711, USA
[4]Marine Sciences, University of Connecticut, Groton, CT, USA
[5]Department of Statistics, University of Connecticut, Storrs, CT, USA

## 1. INTRODUCTION

- The **area of focus is Lake Erie,** the **shallowest, warmest, and most biologically active** of the Great Lakes. The lake provides drinking water for 12 million people in the U.S. and Canada. Agriculture, tourism, commercial fishing, and recreational activities are a few of the ecosystem services that Lake Erie provides but **excessive algal growth** poses threats to the ecosystem and human health. (US Environmental Protection Agency, 2018).
- In this study we aim to **demonstrate how modeled and observed variables** can be used to **identify algal blooms** using chlorophyll-α (chlor-α) concentrations as proxies for the period 2002-2012.
- From 2002-2012 the chlor-α level in the **western** basin has averaged at a eutrophic level while the **central** basin has been mesotrophic and the **eastern** basin oligotrophic (Forage Task Group, 2012).

## 2. SAMPLE LOCATIONS, DATA, AND WATERSHEDS

- chlor-α data was collected by the Lake Erie Committee (LEC) Forage Task Group with stations indicated in green, and the Great Lakes National Program Office (GLNPO) indicated in white (Fig. 1).
- From 2002-2012, samples were taken every two weeks from beginning of April to end of October.
- High chlor-α measurements were identified: a measurement of 80 ug/L and 62 ug/L remained in the data to increase predictability of high chlor-α concentrations (Fig. 2).
- Watersheds were delineated from a HUC 8 scale to determine the drainage area into each sample location.
- Only US watersheds and sample locations were used for this study.

**Geographical Map of Lake Erie**


**Fig. 1: Chlorophyll-α sample stations, and surrounding watershed outlets (in red).** Bars delineate western, central and eastern basin sub-boundaries.

**Chlorophyll-α (ug/L) by Year (2002 – 2012)**

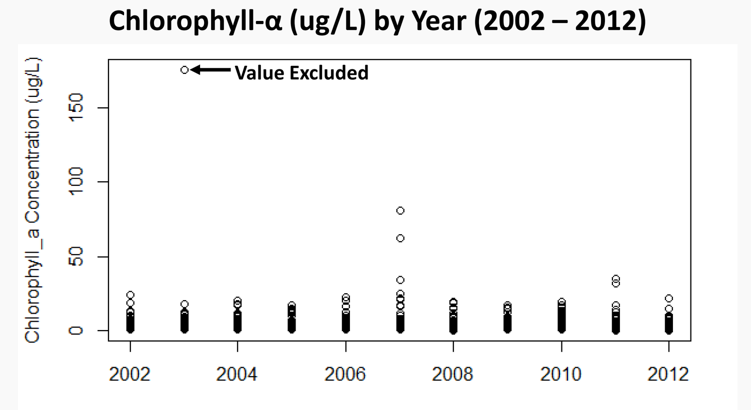


**Fig. 2: Chlorophyll-α (ug/L) plotted by year.** A chlor-α measurement of 175 ug/L was removed from the model due to it being an extreme outlier.

## 3. MODEL DATA

- Consistent inputs of coupled CMAQ-EPIC data (Bash, J. O., E. J. Cooter, et al., 2013) were used alongside WRF and VIC data (Fig. 3).
- Environmental predictors allow understanding of science from source to lake and improve the ability to identify and characterize associations.
- Point model variables (**Point**) were obtained from pairing each sample station to the closest gridded model point.
- Watershed model variables (**WS**) were created from aggregating gridded model points over the watershed area related to each sample station.
- Each model variable was lagged for 5 days resulting in more than 250 predictor variables.

**Model Data Interactions with Lake Water**

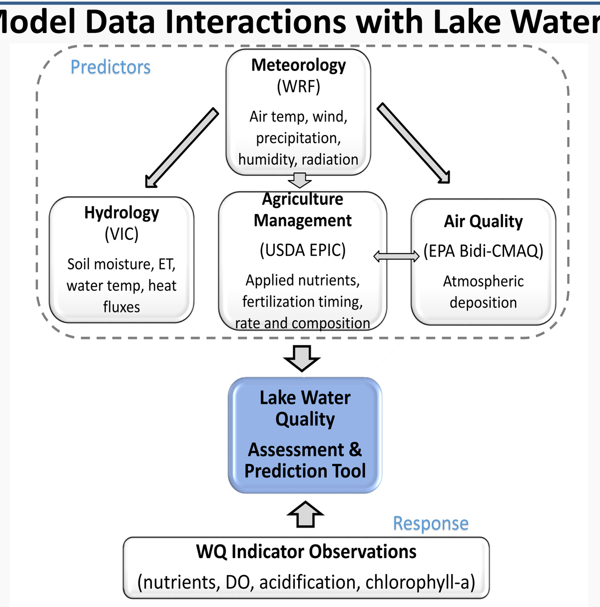


**Fig. 3: Flowchart of model data indicates linked directionality of interactions between modeling systems.**

## 4. MACHINE LEARNING ALGORITHM

**Random Forest (RF) Methodology:**

- Random subsets of model variables are selected at each step and used to create decision trees (Fig. 4). The optimal number of variables to consider for the root node is calculated by the square root of the amount of explanatory variables.
- There are around 250 explanatory variables, therefore, the number of variables tried at each split is 16.
- Each tree gives a classification and saves the trees votes, the forest chooses the classification having the most votes over all the other trees and takes the average of the output by different trees.

**Simplified Random Forest Diagram**

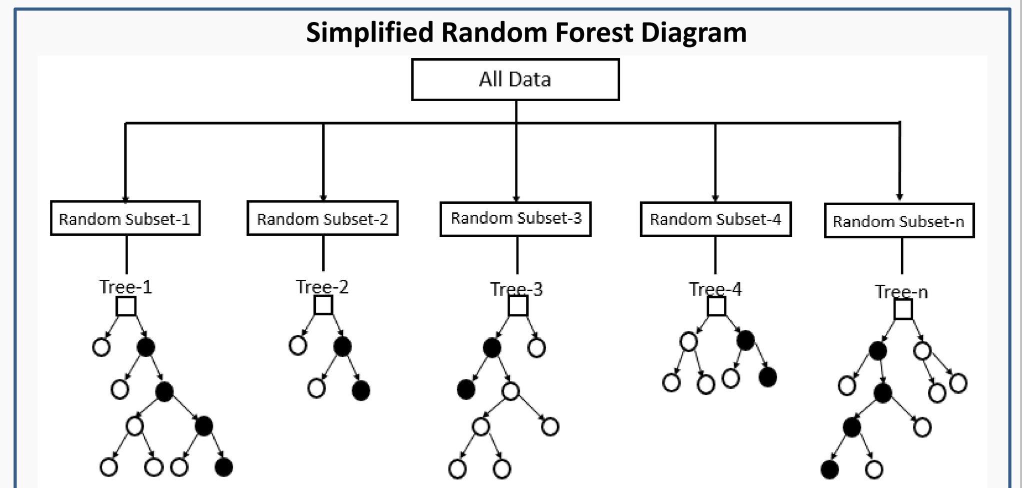


**Fig. 4: Schematic diagram of the random forest model.** Squares indicate root node. Side arrows indicate amount of splits per node. Hollow circles indicate continued node variable runs. Shaded circles indicate decision.

- Raw chlor-α data was used in the RF model.
- No distribution assumption was applied.

## 5. TOP PREDICTORS AND EFFECTS ON chlor-α

**Top Chlorophyll-α Predictors**

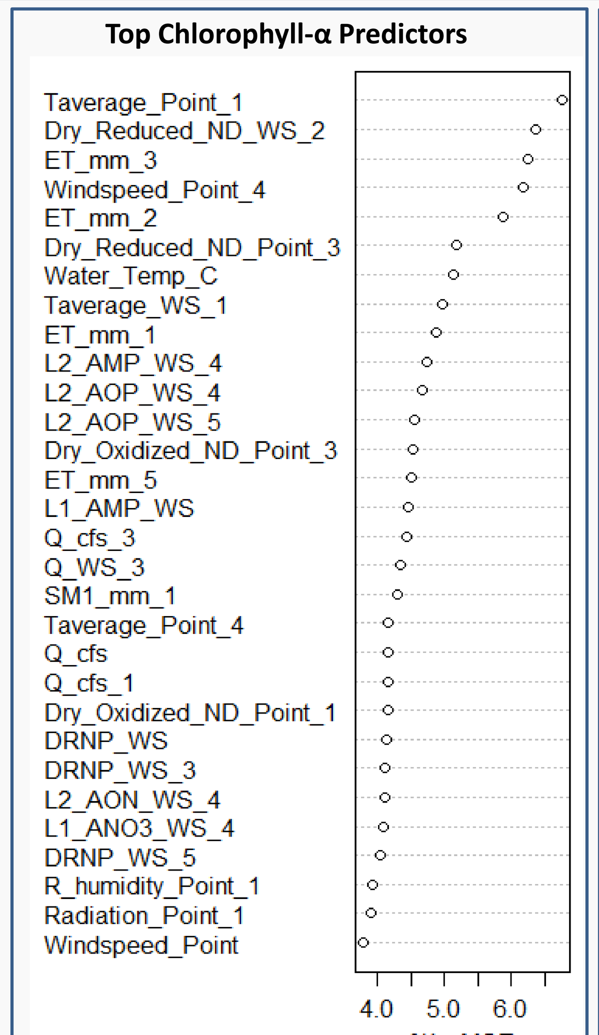


**Fig. 5: Chlorophyll-α variable importance plot.** Number after each variable is the number of lagged days. Point is a grid point model data. WS is an aggregated model data in the watershed.

**Table 1: Definition for top predictors of chlorophyll-α listed in Fig. 5.**

| Top Predictors | Units | Definition | Model |
|---|---|---|---|
| Taverage (Point, WS) | °C | average maximum and minimum air temperature | WRF |
| Dry_Reduced_ND (Point, WS) | kg/ha | dry deposited reduced N | CMAQ |
| ET_mm (Point) | mm | evapotranspiration | VIC |
| Windspeed (Point) | m/s | wind speed | WRF |
| Water_Temp_C (Point) | °C | water temperature | VIC |
| L2_AMP (WS) | kg/ha | 2nd layer mineral phosphorus application rate | EPIC |
| L2_AOP (WS) | kg/ha | 2nd layer organic phosphorus application rate | EPIC |
| Dry_Oxidized_ND (Point) | kg/ha | dry deposited oxidized N | CMAQ |
| L1_AMP (WS) | kg/ha | 1st layer mineral phosphorus applicate rate | EPIC |
| Q_cfs (WS) | cfs | water flow | VIC |
| Q (WS) | mm | runoff | EPIC |
| SM1_mm (Point) | mm | level 1 soil moisture at outlet | VIC |
| DRNP (WS) | kg/ha | soluble phosphorus loss through drainage system | EPIC |
| L2_AON (WS) | kg/ha | 2nd layer organic nitrogen application rate | EPIC |
| L1_ANO3 (WS) | kg/ha | 1st layer nitrate nitrogen application rate | EPIC |
| R_humidity (Point) | | relative humidity | WRF |
| Radiation (Point) | $W/m^2$ | radiation | WRF |

- The out-of-sample cross validation technique applied for the prediction of chlor-α is 10-fold cross validation, repeated 5 times. The overall RF model does a good job predicting chlor-α but under-predicts chlor-α concentrations greater than 30 µg/L (Fig. 6).

**Random Forest Out-Of-Sample Prediction of Chlorophyll-α (ug/L)**

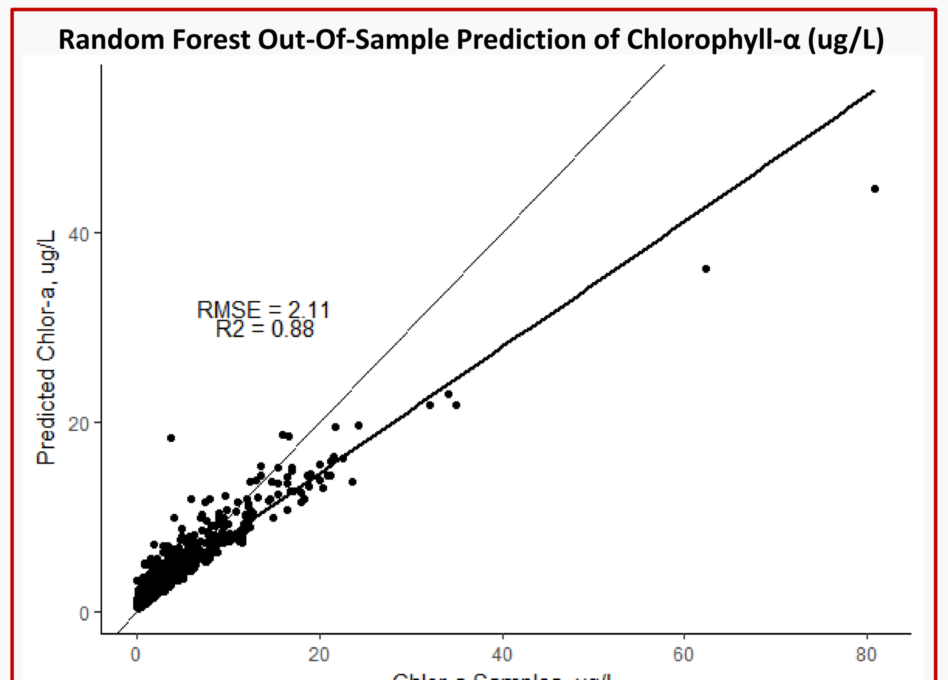


**Fig. 6: Prediction of chlor-α plotted by actual chlor-α samples (ug/L) through random forest. Shown are RMSE and $R^2$. Model underpredicts extreme chlor-α values.**

**Partial Dependence Plots**

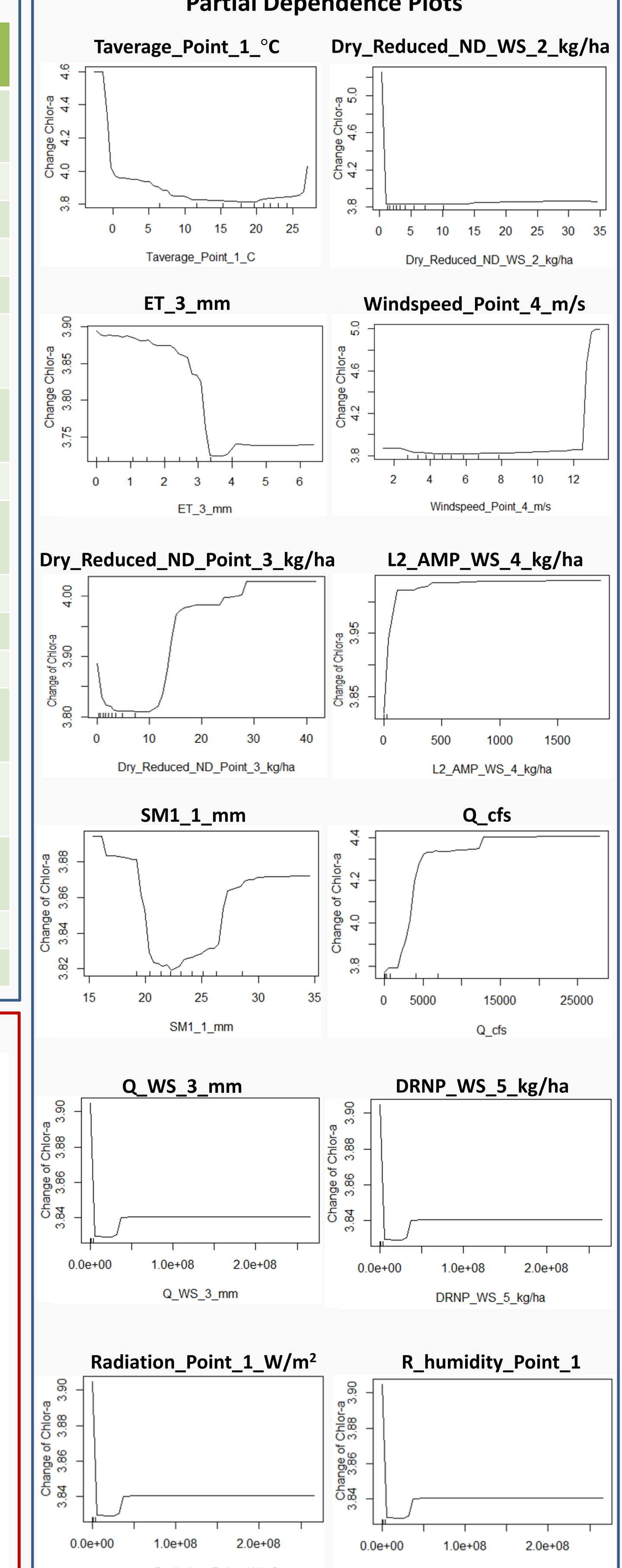


**Fig. 7: Partial dependence plots indicating how chlorophyll-α changes (y-axis) as a consequence of an exploratory variable (x-axis).**

## 6. NEXT STEPS

- This project will be continued as part of Feng Chang's PhD dissertation in Environmental Engineering. A social science component will be added as part of future work.
- A more detailed understanding between the connection of chlor-α and the top environmental variables selected by random forest needs to be established.
- Regression models and other machine learning algorithms will be explored to evaluate and compare the results of random forest to test for similarities.
- The methods applied to the chlor-α data will be tested and applied to predict dissolved oxygen levels, total nitrogen, and total phosphorus data sets in Lake Erie for the years 2002- 2012 with the data provided by the LEC and GLNPO.

## 7. ACKNOWLEDGMENTS

- Data was provided by the Lake Erie Committee Forage Task Group. Special thanks to: James Markham (NY DEC), Patrick Kocovsky (USGS), and Mark Clapsadl (Buffalo State College, Great Lakes Center).
- Data was provided by the Great Lakes National Program Office. Special thanks to: Kenneth Klewin (EPA).
- Part of this work is possible with the support of the U.S. Department of Education's Graduate Assistance in Areas of National Need (GAANN) Fellowship for Christina Feng Chang, Aug 2017 to May 2019.
- Disclaimer: The views expressed in this presentation are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

## 8. REFERENCES

1. US Environmental Protection Agency (2018). "US Action Plan for Lake Erie". https://www.epa.gov/sites/production/files/2018-03/documents/us_dap_final_march_1.pdf. Cited 10/9/2018.
2. Forage Task Group. 2012. Report of the Lake Erie Forage Task Group, March 2012. Presented to the Standing Technical Committee, Lake Erie Committee of the Great Lakes Fishery Commission, Ann Arbor, Michigan, USA.
3. Bash, J. O., E. J. Cooter, et al. (2013). "Evaluation of a regional air-quality model with bidirectional NH3 exchange coupled to an agroecosystem model." Biogeosciences **10**: 1635-1645.
4. US Environmental Protection Agency (2016). "About the Great Lakes National Program Office (GLNPO)". US Environmental Protection Agency, GLNPO, Chicago, IL. https://www.epa.gov/aboutepa/about-great-lakes-national-program-office-glnpo. Cited 10/09/2018.

## CONTACT

Email: Christina.Feng_Chang@uconn.edu
Group Website: http://airmg.uconn.edu.