UCONN
SCHOOL OF ENGINEERING

Assessment of Environmental Variables that Affect Dissolved Oxygen Concentrations in Lake Erie Using Multi-media Modeling and Machine Learning

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

[1]Civil and Environmental Engineering, University of Connecticut, Storrs, CT 06269, USA
[2]National Exposure Research Laboratory, Office of Research and Development, US Environmental Protection Agency, Research Triangle Park, NC 27711, USA
[3]Department of Marine Sciences, University of Connecticut, Groton, CT 06340, USA
[4]Department of Operations and Information Management, University of Connecticut, Stamford, CT 06901, USA
[5]Department of Statistics, University of Connecticut, Storrs, CT 06269, USA

1. INTRODUCTION

- Excessive algal growth in Lake Erie poses threats to the health of the public, the economy, and aquatic organisms. Abundance, competition, death, and decay between algae severely depletes dissolved oxygen (DO) which leads to hypoxic conditions in the system (Chislock et al., 2013).
- Healthy levels of DO are vital for many aquatic organisms. DO levels less than 5 mg/L can kill fish and change fish behavior (*Quality Criteria for Water 1986*, US EPA, 1986).
- In this study, we investigate and predict hypoxia using DO concentrations as proxies for the period 2002-2012 by using modeled and observed variables.
- In the past, we utilized similar techniques presented in this study to understand harmful algal blooms and eutrophication in Lake Erie (Feng Chang et al., 2018).

2. SAMPLE LOCATIONS OF OBSERVED DATA

Geographical Map of Lake Erie

Fig. 1: Dissolved oxygen sample stations (LEC in green; GLNPO in white), and surrounding watershed outlets (in red).

Histogram of DO (mg/L) Data

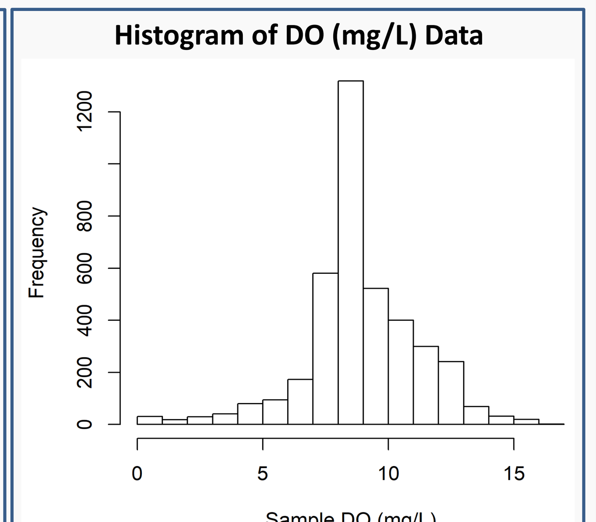

Fig. 2: Frequency distribution of DO at Middle and Bottom layers.

Table 1: Ranges of Surface, Middle, and Bottom layers of DO samples in Western, Central, and Eastern Erie basins.

Depth (m)	Western	Central	Eastern
Surface	0 – 2.4	0 – 3.6	0 – 4.0
Middle	2.5 – 4.7	3.7 – 8.6	4.1 – 12.9
Bottom	4.8 – 10.7	8.7 – 24.4	13 – 48.1

- Focus of this study is in the Middle and Bottom layers where problematic instances of DO <5mg/L tend to occur. A total of 3952 raw DO data points are used (Fig. 2).

3. MODEL DATA

- Consistent inputs of coupled CMAQ-EPIC data (Bash et al., 2013) were used alongside WRF and VIC data (Fig. 3).
- Point variables (**Point**) were obtained by pairing the closest model grid point to each sample station.
- Watershed variables (**WS**) were created by aggregating daily values for all grids in the HUC-8 watershed draining into the lake.
- Static variables quantify the physical characteristics of the lake.
- In addition to each individual variable, each modeled variable was lagged for 5 days resulting in 255 predictor variables.

Model Data Interactions

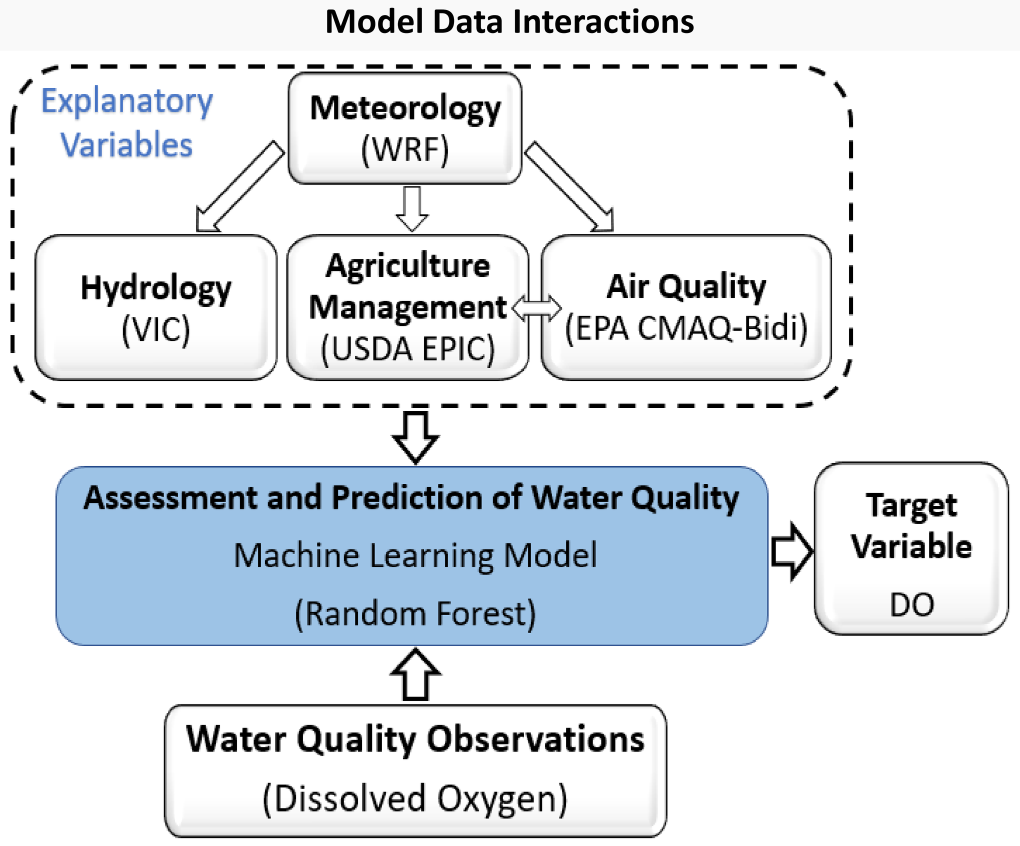

Fig. 3: Flowchart of model data indicating the directionality of interactions between modeling systems.

4. WORK FLOW

- The machine learning (ML) model used is random forest (RF). RF aggregates multiple decision trees to obtain a consensus prediction of the response variable (Fig. 4).
- Step 1: Train and validate RF model with all explanatory variables including a set of generated variables consisting of random numbers used to reduce noise in the model.
- Step 2: Optimize the RF model by tuning hyperparameters: *mtry* and *ntree*.
- Step 3: Examine performance of the RF model through 10-fold cross validation (CV) and evaluate importance of top explanatory variables through accumulated local effect (ALE) plots.

Simplified Random Forest Diagram

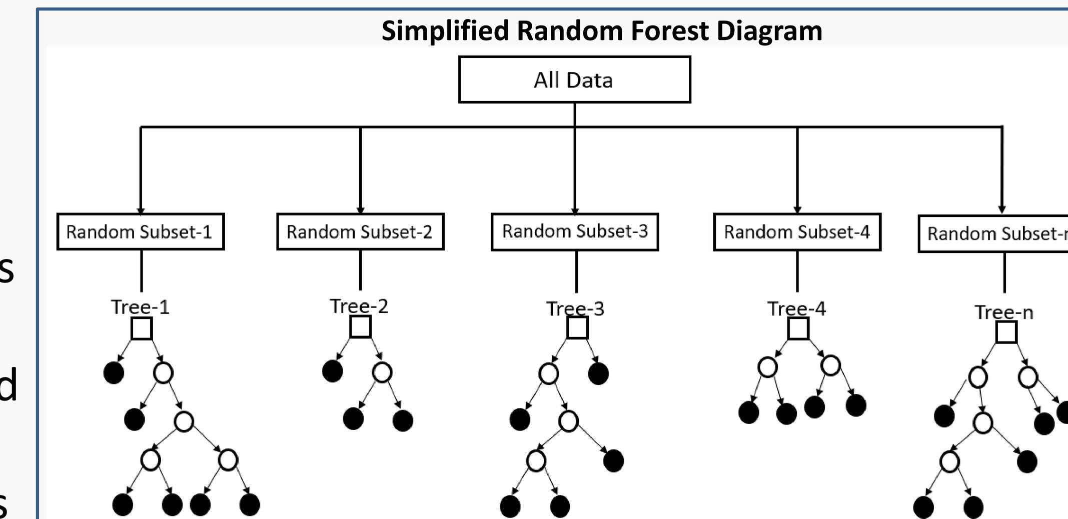

Fig. 4: Simplified diagram of the random forest model. Squares indicate root node. Side arrows indicate a split in the node. Hollow circles indicate continued node splits until the mean square error becomes low (dark circles).

5. TOP PREDICTORS AND EFFECTS ON DO

Top 30 Predictors of DO

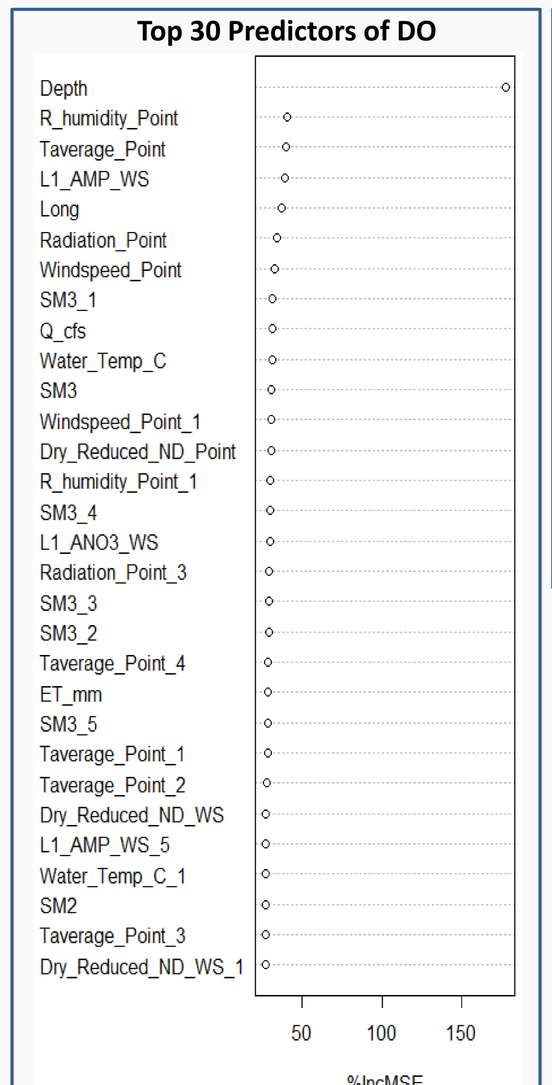

Fig. 5: DO variable importance plot. Higher %IncMSE means higher importance. The number at the end of a variable indicates a lag from 1 to 5 days.

Table 2: Definition of top predictors of DO listed in Fig. 5.

Top Predictors	Units	Definition	Model
Depth	M	in water depth of DO sample in station	
Longitude	degrees (°)	longitude of sample station	
R_humidity_Point		relative humidity	WRF
Radiation_Point	W/m^2	radiation	WRF
Taverage_Point	°C	average air temperature (Tmin+Tmax)/2	WRF
Windspeed_Point	m/s	wind speed	WRF
Dry_Reduced_ND (Point, WS)	kg/ha	dry deposited reduced N	CMAQ
ET_mm	mm	evapotranspiration	VIC
Q_cfs	cfs	waterflow	VIC
SM2	mm	soil moisture layer 2 (10-40 cm)	VIC
SM3	mm	soil moisture layer 3 (40-150 cm)	VIC
Water_Temp_C	°C	water temperature at outlet of closest watershed	VIC
L1_AMP_WS	kg/ha	layer 1 MP (mineralized P) application rate	EPIC
L1_ANO3_WS	kg/ha	layer 1 (N-NO3) nitrate application rate	EPIC

There is an overprediction when DO is less than 5 mg/L (Fig. 6). Overall, 98.3% of the model's predictions are within 2 and 0.5 times the observations.

The model identifies DO < 5mg/L, 48.2% of the time, and DO ≥ 5 mg/L, 99.5% of the time.

The model correctly detects between hypoxic vs. non-hypoxic conditions over 97.0% of the time.

Random Forest Prediction of DO (mg/L)

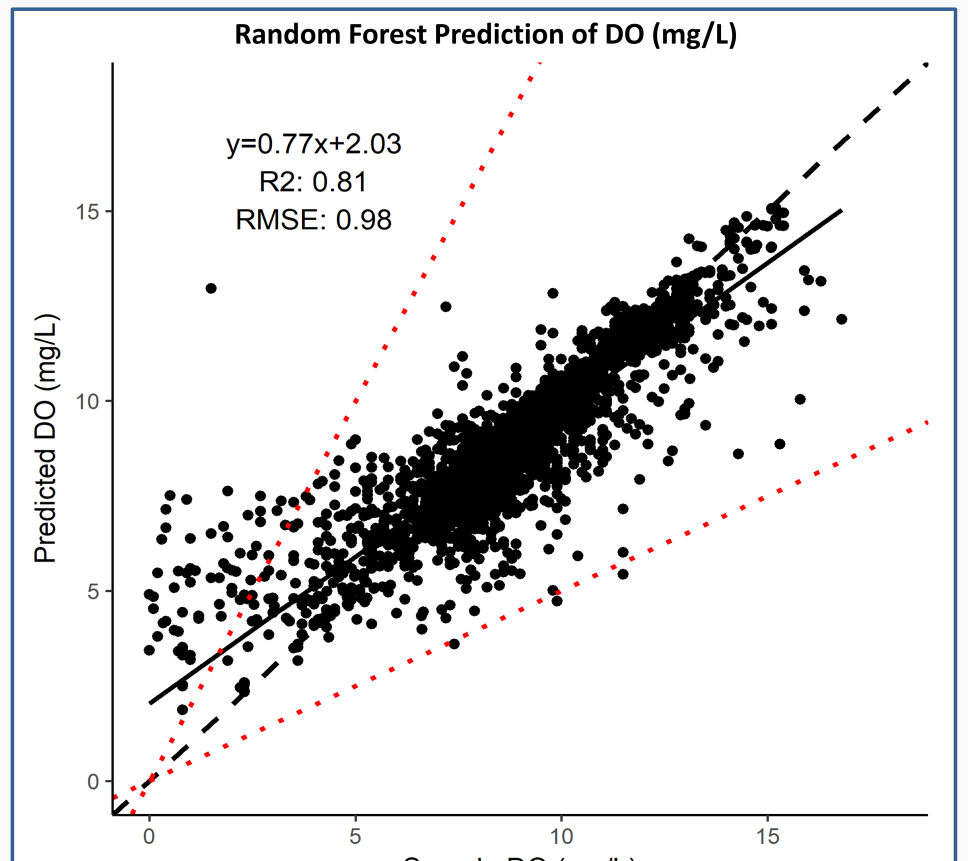

Fig. 6: Prediction of DO at the Middle and Bottom layers. Dashed black line is the one-to-one line; black bold line is the regression line; dotted red lines are the FAC2 lines.

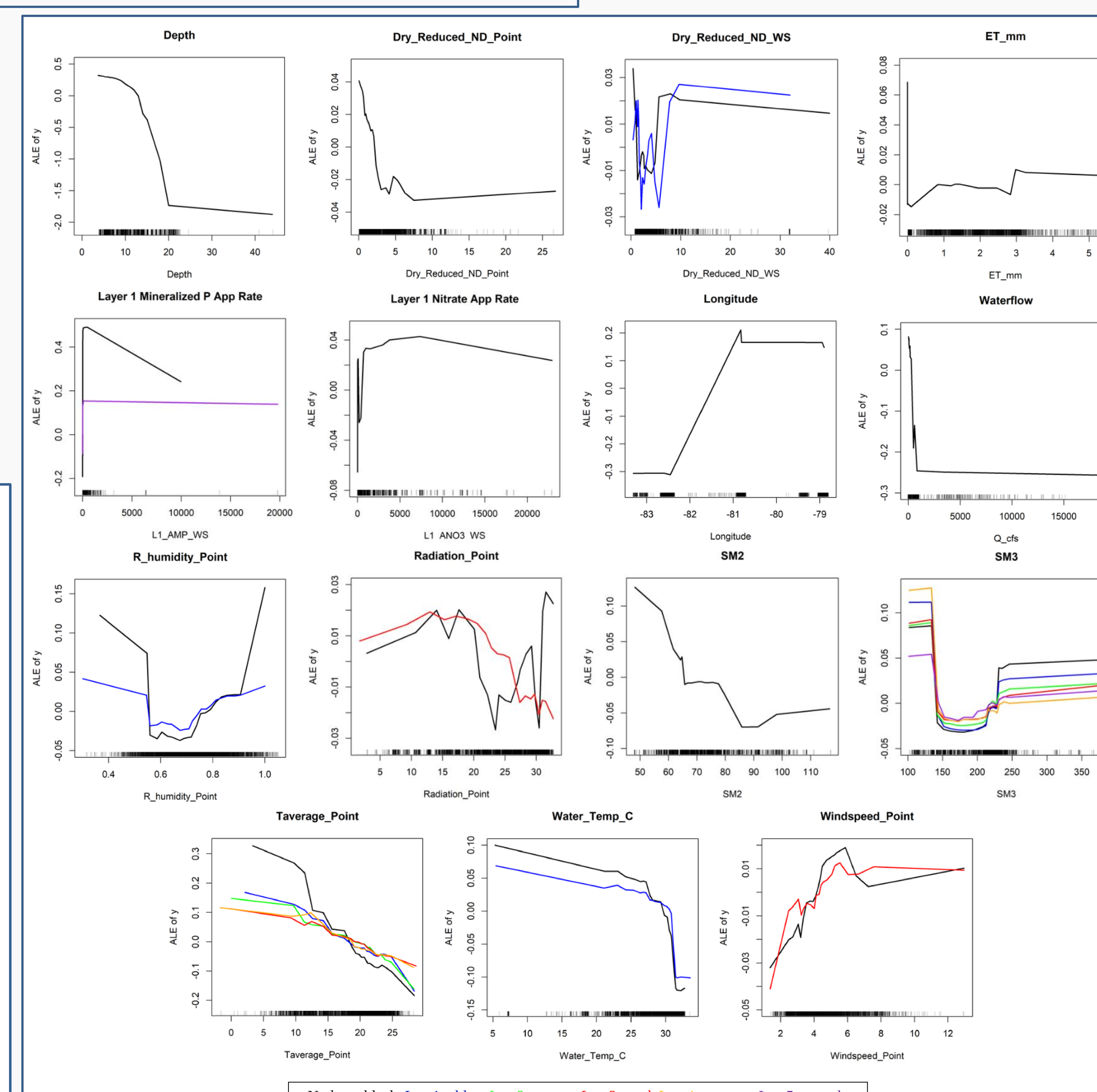

Fig. 7: Overlaid accumulated local effect (ALE) plots for the top variables listed in the variable importance plot in Fig. 5. Each colored line represents a specific lag day according to the legend. ALE plots indicate the change in DO (y-axis) as a consequence of an exploratory variable (x-axis).

6. NEXT STEPS

- A more detailed understanding between the connection of DO and the top environmental predictors selected by the RF model needs to be established.
- Methods applied to the DO data will be tested and applied to predict total nitrogen and total phosphorus data sets in Lake Erie for the years 2002-2012 with the data provided by the LEC and GLNPO.
- Other ML algorithms will be explored to evaluate and compare the results of the RF model.
- The methods applied to Lake Erie can be applied to other Great Lakes, other inland lakes, and coastal locations.

ACKNOWLEDGMENTS

- *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.
- Special thanks to Dr. Ellen Cooter (EPA retiree) who was responsible for conducting the CMAQ-BiDi and EPIC simulations. We would also like to thank James Markham (NY DEC) and Patrick Kocovsky (USGS) from the Lake Erie Committee Forage Task Group for guidance on utilizing LEC data.

REFERENCES

- Bash, J. O.; Cooter, E. J.; Dennis, R. L.; Walker J. T.; Pleim, J. E. Evaluation of a regional air-quality model with bidirectional NH3 exchange coupled to an agroecosystem model. *Biogeosciences*. **2013**. *10*(3): 1635-1645.
- Chislock, M. F.; Doster, E.; Zitomer, R. A.; Wilson, A. E. Eutrophication: causes, consequences, and controls in aquatic Ecosystems. *Nature Education* Knowledge. **2013**. *4*(4): 10.
- Feng Chang, C.; Nowakowski, C.; Astitha, M.; Garcia, V.; Vlahos, P.; Tang, C.; Wanik, D.; Yan, J. Updates on Using Multi-media Modeling to Investigate Conditions Leading to Harmful Algal Blooms. *17th Annual Community Modeling and Analysis System (CMAS) Conference*, University of North Carolina at Chapel Hill's Friday Center, NC, Oct 2018.
- *Quality Criteria for Water 1986*; United States Environmental Protection Agency: Office of Water Regulations and Standards: Washington, DC, 1986.

CONTACT

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