EAS 560 Data Science Project

HW1

Yuan Hui

Project Background and Motivation

Understanding the nutrient loading to lakes is very important to monitoring water quality conditions for the lake, especially for large lakes that provide significant water and ecological resources, such as the Great Lakes (NOAA GLERL 2014). Lake Ontario is the most downstream water body of the Great Lakes. It has the water surface area of 19,000 km^2 and the drainage basin area of 64,000 km^2. (https://www.glerl.noaa.gov/education/ourlakes/lakes.html 2017) It is a critical water resource for the United States as approximately 5.6 million people reside in its watersheds. Therefore, monitoring and controlling nutrient loading to Lake Ontario is important. Another reason to understand the nutrient loading to Lake Ontario is since the mid-1990's, Lake Ontario is facing an acute ecological problem: the widespread resurgence of *Cladophora* at nuisance levels (Dolan and Chapra 2012), as shown in Figure 1(right). Complaints of shoreline fouling, beach closures, and water intake clogging have increased since early 2000's (Dolan and Chapra 2012). Back to 1972 *Cladophora* blooms were one of the issues that prompted development of the Great Lake Water Quality Agreement (GLWQA) between the U.S. and Canada (International Joint Commission 1974). This agreement called for reductions in tributary nutrient loading to the Great Lakes, especially for phosphorus, which is believed to be the limiting nutrient for the growth of *Cladophora* (Dolan and Chapra 2012). Painter and Kamaitis (1987) documented a substantial (58%) decline in *Cladophora* biomass at seven Lake Ontario sites from 1972 (before GLWQA) to 1983 (after GLWQA). Due to this improvement in lake-wide conditions, monitoring *Cladophora* was no longer a priority and this led to a gap in data collection and monitorning between the mid-1980's to late 1990's. These initiatives in 1970's showed that reducing phosphorus loading effectively addressed the *Cladophora* bloom problems and it is believed that in Lake Ontario, phosphorus is the limiting nutrient for *Cladophora* bloom.

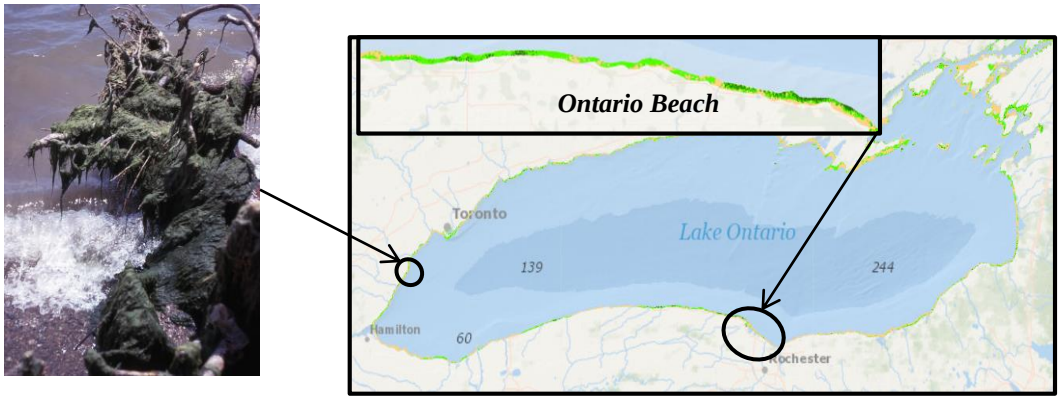

Figure1. (left) *Cladophora* at Jack Darling Park, Mississauga, Sept. 11, 2011 (Photo by W.D. McIlveen), (right) Satellite derived submerged aquatic map for 2010-2011 (*http://geodjango.mtri.org/static/sav/)*

This research is motivated by the need to understand the causes of these ecological issues and to propose potential solutions to reduce their occurrence. As mentioned, phosphorus is believed to be the limiting nutrient for *Cladophora* resurgence, but the phosphorus loading, especially from small tributaries are

unknown due to lack of measurements. Water quality is only monitored in large tributaries (Niagara River, Genesee River and other major rivers), while these data is not enough to understand the phosphorus loading budget to the whole lake. More studies are needed to understand phosphorous loading to the lake.

Project Objective

The objective of this study is to train an Artificial Neural Network (ANN) model with the data from a sub-watershed to estimate phosphorus loading. The trained ANN should have the potential to apply to other sub-watersheds (mainly focus on US. side) to predict phosphorus loading to Lake Ontario. The findings and the model framework from this research could be able to provide decision supporting information for controlling phosphorus loading to the lake through tributaries and watershed.

Methods

The ANN model should be trained and tested for the watershed that has the best datasets. In this study, a calibrated hydrological model has been tested in Genesee River basin and the model is able to provide daily discharge and Total Phosphorus (TP) loading from Genesee River basin to Lake Ontario from January 2008 to December 2017. These TP results are the target of ANN and the input variables involve meteorological data and phosphorus Point Source (PS)/Non-point Source (NPS) inputs.

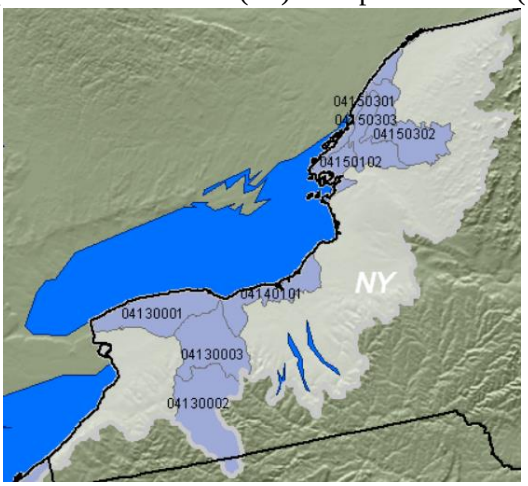

Figure 2. Lake Ontario basin (US side) in grey and Genesee River basin (0413002 and 0413003) in purple.

The target TP is the results from Soil and Water Assessment Tool (SWAT) due to the lake of field water quality measurements. SWAT is one of the well-developed hydrologic quality models to predict the impacts of water management, sediment and agricultural chemical yields in large basins. This model has been applied to Genesee River watershed in Lake Ontario basin, (Makarewicz et al. 2015). The existing model is well-calibrated to reproduce daily TP loading from this sub-watershed. However, building a SWAT model for each sub-watershed in the Lake Ontario basin could be costly in time and computational efforts, and for some small sub-watersheds, there are not enough measurements to calibrate the model, such as watersheds of east branch of TwelveMile Creek, FourMile Creek and Upper Salmon River.

ANN model is introduced to estimate daily TP loading from sub-watersheds. The existing SWAT model for the Genesee River Basin provides the time-dependent TP as target to train and test ANN model. The input layer xi could be the same as the inputs in the SWAT model, as illustrated in Figure 3. The model

hyperparameters should be tested, especially for (1) the ones associated with the model itself, such as the model's number of hidden layers, the number of hidden units, activation functions, weight initialization, random seeds, and data preprocessing, and (2) the ones associated with the gradient- descent algorithm such as learning rate, batch size, number of training iterations, and momentum (Bengio 2012). ANN have been successfully trained to replace SWAT in calculating runoff and sediment yields in other catchments, (Zhang et al. 2009; Singh et al. 2014; Duong et al. 2016; Jimeno-Sáez et al. 2018), so it has the potential to work well in our case.

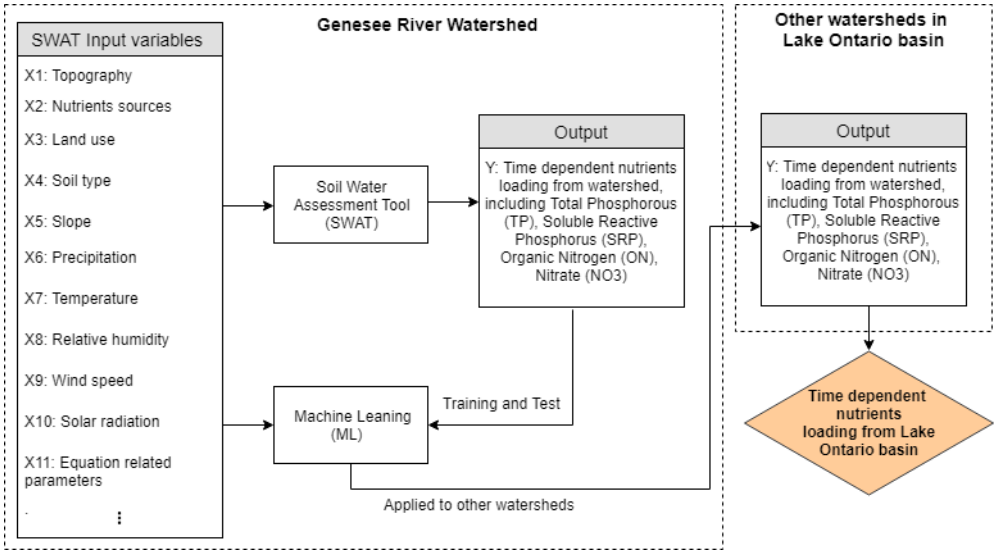

Figure 3. Framework for this project

Data Resource

Target data is daily TP results from SWAT model. Input data includes daily weather data (temperature and precipitation, solar radiation, humidity, evaporation, wind) from the National Weather Service's National Climatic Data Center for the time period 1 January 2008 through 31 December 2017 (NOAA NWS, 2018). Phosphorus related inputs include crop management practices, point sources, confined animal feeding operations, and groundwater phosphorus concentration. Crop rotation and fertilization is based on data provided by the Genesee County Soil and Water Conservation District (George Squires, Genesee County SWCD) and the 2010 Cornell Guide for Integrated Field Crop Management (CCE, 2010). PS inputs of phosphorus are from six Waste Water Treatment Plants (WWTPs): Castile, Perry, Geneseo, Lakeville, Town ofYork, and Avon. Discharge data for the WWTPs were obtained from the USEPA Envirofacts NPDES database (USEPA, 2011). Additional point sources (discharge and phosphorus) include Kodak Park, Morton Salt, and Arkema Inc. based on SPDES permits (USEPA, 2011). 29 confined animal feeding operations (CAFOs) are included as NPS. The amount of manure produced by each CAFO (kg manure/day) was calculated based on the number and type of animals and a constant amount of manure produced per animal per day (ASAE, 2003, 2005).

Expected Outcomes

In the end of this project, the expected outcomes are:

1. The sensitive input variables that affecting watershed TP loading should be identified and analyzed.
2. The ANN model is well trained with the data from Genesee River basin.

3. The model hyperparameters should be tested and the optimal parameters should be selected.
4. Test the model potentials to estimate other high phosphorus loading sub-watersheds