WATER QUALITY AND STREAMFLOW ESTIMATION AT UNGAUGED WATERSHEDS USING MACHINE LEARNING.

G. Mallya\textsuperscript{2}, A. Gupta\textsuperscript{2}, M. M. Hantush\textsuperscript{3}, and R. S. Govindaraju\textsuperscript{1,2}

\textsuperscript{1} Bowen Engineering Head of Civil Engineering & Christopher B. and Susan S. Burke Professor
\textsuperscript{2} Lyles School of Civil Engineering, Purdue University, West Lafayette, IN
\textsuperscript{3} EPA, NRMRL, Cincinnati, OH.

World Environmental and Water Resources Congress
Pittsburg, PA.
May 22, 2019
The findings and conclusions in this presentation have not been formally disseminated by the U.S. EPA and should not be construed to represent any agency determination or policy.
OVERVIEW

1. Introduction
2. Objectives and Study area
3. Methodology
4. Results
5. Conclusions
6. Future Work and Feedback
INTRODUCTION

• Reconstruct past and present water quality for a state or a larger region (e.g. river basin) based on observed water quality data collected at few monitoring stations.
  - Sparse monitoring data (in time, space, and constituents) due to high cost of operation.
  - Continuous water quality monitoring is mostly carried out on streams that are known to be severely impaired.

• Federal Clean Water Act (1972) requires states to provide both
  - Overall health of all streams in a state
  - Identify streams that are impaired (or hotspots), and design TMDLs

• Hydrologic models are popularly used to simulate water quality but require careful model calibration and can be time consuming.
• Water quality reconstruction
  ➢ Often necessary to reconstruct the WQ time series using surrogate variables such as streamflow.
  ➢ Use LOADEST (Runkel et. al., 2004) equations in conjunction with a relevance vector machine (RVM) algorithm.
  ➢ The Bayesian basis of RVM allows reconstruction of model error associated with each reconstructed point.

### INTRODUCTION

**WQ Reconstruction: LOADEST equations (Runkel et. al., 2004):**

<table>
<thead>
<tr>
<th>No.</th>
<th>Regression equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$a_0 + a_1 \ln Q$</td>
</tr>
<tr>
<td>2</td>
<td>$a_0 + a_1 \ln Q + a_2 \ln Q^2$</td>
</tr>
<tr>
<td>3</td>
<td>$a_0 + a_1 \ln Q + a_2 \text{dtime}$</td>
</tr>
<tr>
<td>4</td>
<td>$a_0 + a_1 \ln Q + a_2 \sin(2\pi \text{dtime}) + a_3 \cos(2\pi \text{dtime})$</td>
</tr>
<tr>
<td>5</td>
<td>$a_0 + a_1 \ln Q + a_2 \ln Q^2 + a_3 \text{dtime}$</td>
</tr>
<tr>
<td>6</td>
<td>$a_0 + a_1 \ln Q + a_2 \ln Q^2 + a_3 \sin(2\pi \text{dtime}) + a_4 \cos(2\pi \text{dtime})$</td>
</tr>
<tr>
<td>7</td>
<td>$a_0 + a_1 \ln Q + a_2 \sin(2\pi \text{dtime}) + a_3 \cos(2\pi \text{dtime}) + a_4 \text{dtime}$</td>
</tr>
<tr>
<td>8</td>
<td>$a_0 + a_1 \ln Q + a_2 \ln Q^2 + a_3 \sin(2\pi \text{dtime}) + a_4 \cos(2\pi \text{dtime}) + a_5 \text{dtime}$</td>
</tr>
<tr>
<td>9</td>
<td>$a_0 + a_1 \ln Q + a_2 \ln Q^2 + a_3 \sin(2\pi \text{dtime}) + a_4 \cos(2\pi \text{dtime}) + a_5 \text{dtime} + a_6 \text{dtime}^2$</td>
</tr>
<tr>
<td>10</td>
<td>$a_0 + a_1 \text{per} + a_2 \ln Q + a_3 \ln Q \text{ per}$</td>
</tr>
<tr>
<td>11</td>
<td>$a_0 + a_1 \text{per} + a_2 \ln Q + a_3 \ln Q \text{ per} + a_4 \ln Q^2 + a_5 \ln Q^2 \text{ per}$</td>
</tr>
</tbody>
</table>
OBJECTIVES

• Explore the use of machine learning techniques to predict streamflow and water quality (for different constituents) at ungauged basins.

• Use watershed attributes, long-term climate data, soil and land-use data, fertilizer sales data as explanatory variables in lieu of observed streamflow and water quality.

• Develop spatial maps of water quality loads to aid decision makers in identifying critical source areas.
DATA USED

Study Area
- Upper Mississippi River Basin (UMRB)
- Ohio River Basin (ORB)
- Maumee River Basin (MRB)
2354 HUC-10 basins
(approximate area 550 sq. km)

Daily Streamflow
3071 USGS Streamflow stations

Water quality parameters
3890 EPA STORET stations*
- Suspended sediment concentration (SSC or TSS) – 1859 stations
- Nitrite + Nitrate – 1146 stations
- Orthophosphate – 885 stations

Study period
1966-2014

*214 USGS-NAWQA stations within the study area were also used.
DATA USED

Target variables – (1) Daily USGS Streamflow magnitudes (2) Reconstructed WQ loads at EPA STORET stations.

Explanatory variables (compiled at 3071 USGS stations, 3890 EPA STORET stations, and 2354 HUC-10 basins)

• Precipitation, min. and max. Temperature from NCDC
  ➢ annual, seasonal, monthly, weekly, daily time scales

• Soil properties from STATSGO2 and SSURGO
  ➢ hydrologic soil group percentage, available water storage in top 25 cm

• 2011 NLCD Land use and land cover data

• Drainage area, Watershed slope, Stream order

• Latitude, Longitude

• Fertilizer sales from NASS dataset*

* used for Nitrogen and Phosphorus analysis
The framework described above is for estimating daily streamflow at EPA STORET stations. Same framework is used for WQ load estimation at ungauged HUC-10 basins.
Random Forest Regression Model (Breiman, L., 2001)
- Fits multiple decision trees on different parts of same training dataset leading to reduction in variance
- Can handle thousands of variables and work on large databases
- Addresses the overfitting problem and internally performs cross validation

Stage 1:
Training phase – Data at 70% of USGS stations
Testing phase – Data at remaining 30% of USGS stations
Prediction phase – Streamflow estimates at EPA-STORET stations

Stage 2:
Training phase – Reconstructed WQ Data at 70% of EPA-STORET stations
Testing phase – Reconstructed WQ Data at remaining 30% of EPA-STORET stations
Prediction phase – WQ load estimates at HUC 10 watersheds
Other Machine Learning algorithms tested:

• **Gradient Boosting regressor (Friedman, J., 2001)**
  
  – Uses weak meta-learners e.g. decision trees
  – Decision trees are fit on the negative gradient of the given loss function in an additive manner.
  
  \[ F_{m+1}(x) = F_m(x) + h(x) \]

  where the residual \( h(x) = y - F_m(x) \)

• **AdaBoost regressor (Freund and Schapire, 1995)**
  
  – In adaptive boosting, weak learners are modified at each step to learn from errors of the previous training step. Uses an exponential loss function – more weights on samples with poor fit.
Streamflow - Random Forest Model

Training & Testing phase (70-30 split):
X variables – At each USGS station we include watershed attributes, climate variables etc. available for the period 1966-2014

Y variable – Daily streamflow at each USGS station for the period 1966-2014

Prediction phase:
X variables – watershed attributes, climate variables etc. available for the period 1966-2014 at 3890 EPA STORET stations.

Predict daily streamflow (Y variable) at each EPA STORET station for the period 1966-2014

Variable Importance (15 of 90 variables shown here)
RESULTS

Streamflow – Random Forest Model – Testing phase

Station 4180500

\[ R^2 = 0.94 \]

Station 4180000

\[ R^2 = 0.91 \]
Streamflow – Random Forest Model – Testing phase

USGS 4180500

Flow (cfs) vs. DOY 2000

95% PI

Observed series

DOY 2000
RESULTS

• Streamflow were then estimated at 3890 EPA-STORET stations and 214 USGS-NAWQA stations

• Streamflow were also estimated at ungauged HUC-10 basins (will be used for risk-based TMDL analysis as part of future work)

• Streamflow estimated at EPA-STORET and USGS-NAWQA stations were then used as a surrogate variable within RVM-LOADEST model to reconstruct daily WQ loads.
RESULTS

Total Suspended Sediments (TSS)

- Total Stations = 1859 (UMRB – 1359, ORB – 456, MRB – 44)

- These stations have minimum 30 observations.

- We recorded land use, climate data, soil properties, and watershed attributes for drainage area of each STORET station.

- RVM-LOADEST model was then used to reconstruct TSS-loads using predicted daily streamflow at STORET stations.

- Random forest model was trained on 70% of randomly chosen stations and tested on remaining 30% stations.
RESULTS

Total Suspended Sediments (TSS or SSC) - Random Forest Model – Training phase

Random Forest Regression
Training dataset - 1966 to 2014

\[ R^2 = 0.94 \]

Random Forest Regression
Test dataset

\[ R^2 = 0.82 \]
RESULTS

Total Suspended Sediments (TSS or SSC) – Random Forest - Variable Importance

(15 of 112 variables shown here)
RESULTS

Total Suspended Sediments (TSS or SSC) – Random Forest Model - Testing phase

Station PRI072

$R^2 = 0.61$, 668 data points

Station CCR0001

$R^2 = 0.98$, 55 data points
RESULTS

Total Suspended Sediments (TSS or SSC) – Random Forest Model - Testing phase

Station CCR0001 : Reconstructed SSC

<table>
<thead>
<tr>
<th>DOY 2000</th>
<th>95% PI</th>
<th>Median</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Station PRI072 : Reconstructed SSC

<table>
<thead>
<tr>
<th>DOY 2000</th>
<th>95% PI</th>
<th>Median</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>150</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>250</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>350</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
RESULTS

Total Suspended Sediments (TSS or SSC) – Average daily load

Year: 2005

Year: 2010

Units for load is $\text{kg/day/year}$ in log scale.
CONCLUSIONS

• The study showed an application of machine learning techniques to predict streamflow at ungauged EPA-STORET stations.
• Streamflow predictions were then used as a surrogate to reconstruct water quality time series using RVM-LOADEST methodology at EPA-STORET stations.
• Reconstructed WQ loads (SSC, Nitrogen, etc.) at EPA-STORET stations and watershed attributes were used to obtain load estimates at ungauged HUC-10 basin.
• Robust estimates of water quality loads were obtained when we had sufficient training data.
• Spatial maps of water quality loads were developed to help decision makers in identifying critical source areas.
FUTURE WORK

• Compare the results of machine learning models with those obtained from other models (e.g. SPARROW) for different water quality constituents.

• An ensemble of multiple ML models will be used to obtain final estimates of streamflow and water quality loads.

• Use the predicted streamflow and water quality loads at ungauged basins for:
  ➢ Computing reliability, resilience, vulnerability, and watershed health metrics.
  ➢ Performing risk-based TMDL analysis.

• Use ML model predictions to identify HUC-10 watersheds that are ‘hotspots’ for sediment and nutrient loadings and then extend the methodology to estimate the timing and spatial extent of harmful algal blooms in Lake Erie.
THANK YOU

The floor is open for questions ...
USGS-NAWQA DATA USED

Water quality parameters
214 USGS NAWQA stations

- Suspended sediment concentration (SSC) – 151 stations
- Nitrite + Nitrate – 70 stations
- Orthophosphate – 49 stations

Study period
1966-2014
Station CCR0001 : Reconstructed SSC

log(Load in kg/day)

DOY 2000

GB
AB
RF
Observations
RESULTS

Nitrogen (Nitrate + Nitrite)

- Stations = 1146 (UMRB – 857, ORB – 278, MRB – 11)

- These stations have minimum 30 observations.

- We recorded land use, climate data, soil properties, fertilizer sales, and watershed attributes for drainage area of each STORET station.

- RVM-LOADEST model was then used to reconstruct Nitrogen-loads using predicted daily streamflow at STORET stations.

- Random forest model was trained on 70% of randomly chosen stations and tested on remaining 30% stations.
RESULTS

Nitrogen (Nitrate + Nitrite) – Random Forest Model

Training Set (70% stations)  \[ R^2 = 0.92 \]

Test Set (30% stations)  \[ R^2 = 0.87 \]
RESULTS

Nitrogen (Nitrate + Nitrite) – Random Forest Model - Variable Importance

AreaKM2
StreamOrde
LU_Agri
PRCP 3D
FertSales
SUM_AWS025
PRCP 1D
HYDGRP_B
HYDGRP_A
PRCP 7D
LU_Forest
LU_Urban
Longitude
HYDGRP_C
WS_Slope

Importance

(15 of 113 variables shown here)
RESULTS

Nitrogen (Nitrate + Nitrite) – Random Forest Model - Testing phase

Station BSW005

$R^2 = 0.94$, 302 data points

Station K9900000

$R^2 = 0.72$, 240 data points
RESULTS

Nitrogen (Nitrate + Nitrite) – Average daily load

Units for load is kg/day/year in log scale