Use of Surrogate Regressions for Computing Continuous Loads for the USGS Great Lakes Tributary Monitoring Network

Dale Robertson, Laura Hubbard, David Lorenz\textsuperscript{2}, Laura DeCicco\textsuperscript{3}, and Daniel Sullivan

U.S. Geological Survey
Wisconsin WSC, \textsuperscript{2}Minnesota WSC, \textsuperscript{3}OWI

April 6, 2017
*dzrobert@usgs.gov
(608) 821-3867
**GLRI Action Plan II**

*GLRI Action Plan II* summarizes the actions that federal agencies plan to implement during FY15-19 using Great Lakes Restoration Initiative funding — actions to protect and restore the largest fresh surface water system in the world. These actions will build on restoration and protection work carried out under the first GLRI Action Plan, with a major focus on:

- Cleaning up Great Lakes Areas of Concern
- Preventing and controlling invasive species
- Reducing nutrient runoff that contributes to harmful/nuisance algal blooms
- Restoring habitat to protect native species
Monitored areas represent 31-57% of the US drainages.
Flow

Automated samplers
Suspected sediment
Nutrients:
  TP, OP
  TN, NO$_{2+3}$

Multiparameter sondes (surrogates)
YSI-6920’s > EXO’s
Dissolved oxygen
Specific conductance
pH
Temperature
Turbidity
Estimating Loads at the Monitoring Sites – with Regression Approach (General Forms Possible)

Load = f (flow, seasonality, w/ and w/o surrogates)

Form of the **Daily** Load Flow Models (daily estimates) Traditional - Model 1

\[
\ln(\text{Load}_D) = a \ln(Q_D) + b \ln(Q^2_D) + c \sin(j\text{day}) + d \cos(j\text{day}) + e
\]

Form of the **UV** Load Flow Models (every 5 minutes) Model 2

\[
\ln(\text{Load}_{UV}) = a \ln(Q_{UV}) + b \ln(Q^2_{UV}) + c \sin(j\text{day}) + d \cos(j\text{day}) + e
\]

Form of the Load **Surrogate** Models (every 5 minutes) Model 3

\[
\ln(\text{Load}_{UV}) = a \ln(Q_{UV}) + b \ln(Q^2_{UV}) + c \sin(j\text{day}) + d \cos(j\text{day}) + e(\text{Surrogate}_1) + f(\text{Surrogate}_2) + \ldots
\]
Load Computation Process

**Data Pre-Processing**
- Data: NWIS web via R dataRetrieval package
- Fill in missing Unit Value flow data (e.g. ice) with daily flow
- Correct Surrogate data (turbidity/conductance)

**Flow & Surrogate Model Development**
- Choose Models
- Model using flow and seasonality (Traditional Approach) (Model 1)
- Using UV data (Model 2)
- Model including flow & surrogates (Model 3)

**Calibrate, Evaluate the models & Compute Loads**
Estimating Loads at the Monitoring Sites – with Regression Approach

(Determining Specific Form of Model)

Load = f (flow, seasonality, w/ and w/o surrogates)

Form of the Daily Load Flow Models (daily estimates) Traditional - Model 1

\[ \ln(\text{Load}_D) = a \ln(Q_D) + b \ln(Q_D^2) + c \sin(jday) + d \cos(jday) + e \]

Form of the UV Load Flow Models (every 5 minutes) Model 2

\[ \ln(\text{Load}_{UV}) = a \ln(Q_{UV}) + b \ln(Q_{UV}^2) + c \sin(jday) + d \cos(jday) + e \]

Form of the Load Surrogate Models (every 5 minutes) Model 3

\[ \ln(\text{Load}_{UV}) = a \ln(Q_{UV}) + b \ln(Q_{UV}^2) + c \sin(jday) + d \cos(jday) + e(Surrogate_1) + f(Surrogate_2) + \ldots \]

Only evaluated with water quality collected during positive flows
Variable Selection and Order to add into model – Total Phosphorus

<table>
<thead>
<tr>
<th>River</th>
<th>Best UV-flow Models</th>
<th>Variables Included in Surrogate Models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q, Q, Q^2, sin, cos</td>
<td>Q^2, ln (DO), ln (Turb), SC, pH, Temp.</td>
</tr>
<tr>
<td>St. Louis</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Bad</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Ontonagon</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Ford</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Menominee</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Fox</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Manitowoc</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Milwaukee</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Indiana Canal</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Burns Ditch</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>St. Joseph</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Kalamazoo</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Grand</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Muskegon</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>AuSable</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Rifle</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Saginaw</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Clinton</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Rouge</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Huron</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Raisin</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Maumee</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Portage</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Vermilion</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Black</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Cuyahoga</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Cattaraugus</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Genesee</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Oswego</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>St. Regis</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

Count Summary for variable type:

- Q: 16
- Q^2: 12
- ln (DO): 10
- ln (Turb): 13
- SC: 4
- pH: 21
- Temp: 9
- Q, sin, cos: 10
- Q, Q^2, sin, cos: 11
- ln (DO), ln (Turb), SC, pH, Temp: 12

- Total: 3
- Q: 18
- Q^2: 25
- ln (DO), ln (Turb), SC, pH, Temp: 19
- Q, sin, cos: 3
- Q, Q^2, sin, cos: 1
Variable Selection for TP – Point of diminishing Return

Final Models:
- Model 1 & 2: Flow and Seasonality
- Model 3: Flow, Seasonality, and Ln(Turbidity)

Model 2 – Using flow and seasonality

Final Models:
- Model 1 & 2: Flow and Seasonality
- Model 3: Flow, Seasonality, and Ln(Turbidity)
Load Computation Process

Data Pre-Processing
- Data: NWIS web via R dataRetrieval package
- Fill in missing Unit Value flow data (e.g. ice) with daily flow
- Correct zero turbidity/conductance

Surrogate & Flow Model Development
- Choose Models
- Model using flow and seasonality (Traditional Approach) (Model 1)
- Using UV data (Model 2)
- Model including surrogates (Model 3)

Calibrate, Evaluate the models & Compute Loads
- Adjust Negative Flows
- Calibrate Models 1, 2, and 3
- Compute Loads Final “hybrid” Surrogate load
Load Calculation:
Adjust for negative flows – Conserve Mass

Method conserves mass – does not neglect negative flows, it borrows flow from nearest neighbors
Final Load Models to estimate Total Phosphorus Loads (Calibrating the Models)

Form of the Daily Load Flow Models (daily estimates) Traditional - Model 1

\[ \ln(\text{Load}_D) = a \ln(Q_D) + b \sin(j\text{day}) + c \cos(j\text{day}) + d \]

Form of the UV Load Flow Models (every 5 minutes) Model 2

\[ \ln(\text{Load}_{UV}) = a \ln(Q_{UV}) + b \sin(j\text{day}) + c \cos(j\text{day}) + d \]

Form of the Load Surrogate Models (every 5 minutes) Model 3

\[ \ln(\text{Load}_{UV}) = a \ln(Q_{UV}) + b \sin(j\text{day}) + c \cos(j\text{day}) + e(\ln(\text{turb})) + f \]

Only calibrated with water quality collected during original positive flows
Total Phosphorus Loads – Final Models

- Daily Flow Model
- UV Flow Model
- Surrogate Model
How much variability in loads/concentrations do the models explain?

Variation Explained by Flow and Surrogate Models

% Variance in Conc. Explained
Mean – 61%
Median – 65%

Improvement
Mean – 33%
Median – 40%
Differences in Daily/Subdaily Load Estimates

Fox River, Wis.

Total Phosphorus Concentrations
Differences in Monthly/Annual Load Estimates

Phosphorus Loads in the Fox River

<table>
<thead>
<tr>
<th></th>
<th>Loadest Daily</th>
<th>Monthly</th>
<th>Annual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loadest UV</td>
<td>5.15%</td>
<td>-1.6%</td>
<td></td>
</tr>
<tr>
<td>UV Surrogate</td>
<td>15.41%</td>
<td></td>
<td>-7.3%</td>
</tr>
</tbody>
</table>
Maumee River at Waterville

Monthly TP Load Comparisons

Comparison with Loadest Daily | Monthly | Annual
---|---|---
Loadest UV | 0.8% | 0.1%
UV Surrogate | 12.0% | -3.2%
Maumee River at Waterville

Compared to Heidelberg – Monthly Difference

WRTDS run for 1975-2014 using USGS data

<table>
<thead>
<tr>
<th>Comparison with Heidelberg - Beales</th>
<th>Monthly</th>
<th>Annual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loadest Daily</td>
<td>30.4%</td>
<td>21.0%</td>
</tr>
<tr>
<td>UV Surrogate</td>
<td>18.7%</td>
<td>17.1%</td>
</tr>
<tr>
<td>WRTDS</td>
<td>21.4%</td>
<td>12.5%</td>
</tr>
</tbody>
</table>
Heidelberg’s Data from the same days are about 15% less than that collected by the USGS.
But how much improvement is there by using unit value flows and by using surrogates??

Compare Measured versus Predicted unit value loads and compute RMSE, Nash-Sutcliffe index, and Average Bias
Comparison of Model 1 and Model 3 based on RMSE
Comparison of Standard Flow Models with Surrogate Models Observed vs Predicted Loads

Median Improvement
RMSE (15-27%)

Nash-Sutcliffe (7-20%)

Bias (56-70%)
Average annual yield (kg/km²/yr) of (A) total phosphorus, (B) total nitrogen, and (C) total suspended sediment for water years 2012-13.
The data used to produce this plot are provisional and have not been reviewed or edited. They may be subject to change.

**EXPLANATION**
- Discharge
- Measured or computed water-quality constituent
- 90-percent prediction interval for computed value
- Value obtained from discrete sampling and analysis
- Value calculated using laboratory analysis and discharge
- Water-quality criteria
Questions?

Contact Information:
Dale Robertson
dzrobert@usgs.gov
608-821-3867
Eutrophication Issues in the Great Lakes

Driven by excess nutrient inputs (primarily phosphorus)

Pelee Island, Lake Erie

Cladophora on Beaches
GLRI Action Plan II

GLRI Action Plan II summarizes the actions that federal agencies plan to implement during FY15-19 using Great Lakes Restoration Initiative funding — actions to protect and restore the largest fresh surface water system in the world. These actions will build on restoration and protection work carried out under the first GLRI Action Plan, with a major focus on:

- Cleaning up Great Lakes Areas of Concern
- Preventing and controlling invasive species
- Reducing nutrient runoff that contributes to harmful/nuisance algal blooms
- Restoring habitat to protect native species
Which approach should give the best load estimates? Comparison of Daily Flow Model Vs UV Flow Model and UV Surrogate Model Observed vs Predicted Loads
Conclusions

1. The monitoring program estimates loads from 31-57% of the U.S. area draining to the Great Lakes.

2. Inclusion of surrogates into regression models:
   A. improves the assessment of short-term daily and monthly loads.
   B. should enable us to assess long-term changes in loads to be more easily detected.

3. By combining surrogate regressions with traditional regressions, continuous loads (with confidence limits) were able to be computed.
Where do we go from here?

A. Publish approach and load estimates

B. Present real-time loads on the internet
Monitoring Design

- Monthly fixed-interval samples
- 6 storm events per year, several samples per event