Journal of Great Lakes Research 44 (2018) 26–42

Contents lists available at ScienceDirect

Journal of Great Lakes Research

journal homepage: www.elsevier.com/locate/jglr

A surrogate regression approach for computing continuous loads for the tributary nutrient and sediment monitoring program on the Great Lakes

Dale M. Robertson [a,*], Laura E. Hubbard [a], David L. Lorenz [b], Daniel J. Sullivan [a]

[a] U.S. Geological Survey, Wisconsin Water Science Center, 8505 Research Way, Middleton, WI 53562, United States
[b] U.S. Geological Survey, Minnesota Water Science Center, 21280 Woodale Drive, Mounds View, MN 55112, United States

ARTICLE INFO

Article history:
Received 7 July 2017
Accepted 4 October 2017
Available online 5 January 2018

Communicated by Craig Stow

Keywords:
Loading
Nutrients
Sediment
Regression

ABSTRACT

Water quality (WQ) in many Great Lake tributaries has been degraded (increased nutrient and sediment concentrations) due to changes in their watersheds, resulting in downstream eutrophication. As part of the Great Lakes Water Quality Agreement, specific goals were established for loading of specific constituents (e.g., phosphorus). In 2010, the Great Lakes Restoration Initiative was launched to identify problem areas, accelerate restoration efforts, and track their progress. In 2011, the U.S. Geological Survey established a monitoring program on 30 tributaries to the lakes, representing ~46% of the U.S. draining area and the spectrum of land uses. Discrete measurements of nutrients and suspended sediment, and continuous measurements of flow and WQ surrogates (turbidity, temperature, specific conductance, pH, and dissolved oxygen) are being collected in these tributaries to document their WQ and estimate continuous (5-min) loading. To estimate loadings, two regression models were developed for each constituent for each site: one using continuous flow and a seasonality factor; and one using flow, seasonality, and continuous surrogates. Variables included in the final models for each constituent were chosen from the explanatory variables that worked "best" for all sites. In computing loads, when continuous surrogate data were unavailable for short periods, loads were computed using the flow and seasonality models. Prediction intervals for all loads were calculated using results from both models. These results provide a better understanding of short-term variability and long-term changes in loading affecting the environmental health of the Great Lakes than traditional regression techniques that employ only flow and seasonality parameters.

Introduction

The Laurentian Great Lakes receive nutrients and sediment from many tributaries draining areas with widely varying land cover ranging from pristine forests to intensive agriculture to large urban centers. Nutrient delivery from these tributaries is extremely variable (Robertson and Saad, 2011). This nutrient loading (mass of a constituent passing a point over a specified time period) has caused eutrophication to various degrees and geographic scales. One of the most noticeable signs of eutrophication is excessive algae growth often leading to harmful algal blooms (Zhou et al., 2015) that may restrict recreational uses and lead to drinking water problems, such as experienced in Toledo, OH in 2014 (New York Times, 2014). Suspended sediment and siltation are the most common stressors affecting streams throughout the U.S. (USEPA, 1998) that can reduce stream clarity and affect sight-feeding fish, and interfere with water-treatment processes, restrict recreational uses and decrease water depth.

In an effort to improve the water quality of the Great Lakes, U.S. and Canada signed the Great Lakes Water Quality Agreement (GLWQA) in 1972, which was renewed in 1978 and identified phosphorus (P) as the nutrient of primary concern and defined target P loads for each lake. The GLWQA was amended in 1987 and updated in 2012 to ensure the chemical, physical, and biological integrity of the Great Lakes (USEPA, 2016). Annex 4 of the GLWQA states that U.S. and Canada are committed to coordinating binational actions to manage P concentrations and loadings, and other nutrients if warranted, in the waters of the Great Lakes. In 2010, the Great Lakes Restoration Initiative (GLRI) was implemented to: target the most significant problems identified in the GLWQA including reducing nutrient and sediment loading; accelerate restoration efforts; and quantitatively track the progress made in addressing the problems (USEPA, 2015). Therefore, it is important to be able to accurately estimate loads of various constituents from selected monitored tributaries and to the lakes as a whole.

Monitoring of Great Lake tributaries has changed dramatically through time. Following the signing of the GLWQA in 1972, monitoring of nutrients, sediment, and major ions was started by most states and provinces around the Great Lakes and by the U.S. Geological Survey (USGS) National Stream Quality Accounting Network (NASQAN). Monitoring programs have been extremely variable both with respect to sampling protocols (continuous versus rotational) and sampling frequency (3 to 400 samples per year) because of different program

* Corresponding author.
E-mail address: dzrobert@usgs.gov (D.M. Robertson).

https://doi.org/10.1016/j.jglr.2017.10.003

objectives (Robertson, 2003). As part of the NASQAN program, >40 tributaries began being sampled with specific sampling protocols in 1974, collecting samples at uniform time intervals without concern for hydrologic patterns of high or low flow (USGS, 2015). The number of sites and sampling frequency gradually decreased through time (from monthly to quarterly); and, by 1995, no sites remained in the Great Lakes Basin. A consistent monitoring program of Great Lakes tributaries would help in properly quantifying the mass fluxes or loads of constituents, and trends in the loads and concentrations.

To estimate loads, time series of concentrations of specific constituents are combined with time series of flow. Continuous flows are now collected at most detailed monitoring sites. However, continuous concentrations (daily or finer resolution) are rarely available in Great Lakes tributaries, with the exception of a few tributaries to Lake Erie monitored by Heidelberg University. Two general approaches have been developed to estimate loads in rivers when discrete data are insufficient to adequately describe the daily temporal variation in concentrations: ratio-estimator methods and regression (rating-curve) methods. The stratified (Beale) ratio-estimator method (Beale, 1962; Cochran, 1977; Dolan et al., 1981) has been the most commonly used approach to estimate annual loading to the Great Lakes (Dolan and Chapra, 2012; Maccoux et al., 2016). In this approach, discrete daily loads are computed annually for each day a site is sampled assuming the discrete concentrations represent the daily average, and then these data are subdivided into one or more strata depending on the nature of flow and concentration relationships. These relationships are then used to estimate the total load for all days in each specified stratum using the ratio of unsampled flow to the total flow in each stratum. Total annual loads are then estimated by summing the loads for each stratum and prediction intervals for the loads are estimated based on the variance in each stratum. The accuracy of ratio-estimator annual loads improves substantially as the number of days sampled increases. The ratio-estimator has been termed unbiased because the mean of several ratio estimates tends toward the "true" mean (Dolan et al., 1981). However, with limited data collected in a given year, such as that now being collected in most tributaries, it is difficult to define different strata and accurately quantify the load and corresponding prediction intervals. Cochran (1977, p. 153) stated that the variance estimate associated with ratio estimators is only reliable if the sample size in each stratum exceeds 30 and is large enough that the coefficients of variation of mean discharge and load are both <10%.

With regression methods, daily loads for all monitored and unmonitored periods are typically estimated using a relation developed between concentrations (or loads) and daily average flows (and other independent variables) measured over several years. This method began as simple linear relations between co-occurring measured concentrations (or loads) and flows but has been modified to account for nonlinearities, seasonal and long-term variability, censored data, biases associated with using logarithmic transformations, and serial correlations in the residuals of the analyses (Cohn, 1995). Typically, only variables describing flow and seasonality (sine and cosine of the day of the year) are included in the regressions. The regression method is often used with sparsely collected data assembled over several years, which may be insufficient for estimating accurate annual loads using other techniques, such as the ratio-estimator technique. With the regression method, daily average flows are typically used to estimate total daily loads because this is the resolution of flows most readily available. Both approaches are usually considered "big river" approaches because they typically are based on the assumption that concentrations from discrete samples are representative of the daily average and changes in concentration (or loads) are estimated on a daily time step. However, these approaches are commonly used to estimate loads in small streams with concentrations changing at time scales less than a day (Walker, 1996).

The ratio-estimator and regression techniques have been used to estimate loads to the Great Lakes using daily data. The ratio-estimator approach was used to estimate nutrient and sediment loadings during 1975–76 when detailed data were collected (Sonzogni et al., 1978). This approach has been continued to be used, even as the temporal frequency of monitoring decreased (Dolan and McGunagle, 2005; Dolan and Chapra, 2012; Maccoux et al., 2016). This approach is used by Heidelberg University (2016) to estimate annual loads from sites sampled almost daily and more frequently during high flows. Because of the cost of detailed (daily discrete) sampling of water-quality constituents, there is a need to find a valid approach to estimate accurate loads without the cost of collecting and analyzing high frequency samples.

The regression approach has been used to estimate nutrient and sediment loads into Lakes Michigan and Superior (Robertson, 1997), Ontario (Hayhurst et al., 2016), and Saginaw Bay (Tao et al., 2010). Robertson and Saad (2011) also used a regression approach to estimate long-term average nutrient loads at sites throughout the entire Basin, which were used to calibrate the SPAtially Referenced Regressions On Watershed attributes (SPARROW) models used to estimate total loading to the Great Lakes. Park et al. (2015) developed a web-based tool to estimate pollutant loading using regression techniques and daily data.

Although flow is often estimated on a continual basis (i.e., every 5–15 min), regression models for load estimation typically have been developed using daily flows because those data are more readily available. Recently, the regression approach has been modified to estimate loads (and prediction intervals) at a finer temporal resolution (Rasmussen et al., 2009). High resolution models (Jones et al., 2011) also often include continuous in-situ measurements, such as turbidity, specific conductivity, and water temperature, which may be correlated with specific water-quality (WQ) constituents, and therefore are referred to as WQ surrogates. The continuous load estimates are then summarized into daily and longer term loads. Densmore et al. (2016) describe techniques to use standard errors of high resolution models to obtain prediction intervals for the daily estimates.

Regression techniques have been used to estimate high-resolution (continuous) loads when all input variables (surrogates) are available (Rasmussen et al., 2009; Baldwin et al., 2012; Densmore et al., 2016); however, continuous surrogate data are often unavailable due to equipment failure, making long-term continuous load estimation difficult. One approach to overcome this difficulty is to develop two regression equations: one based on flow, seasonality, and surrogates and one based only on flow and seasonality. This approach enables continuous loads even when surrogate data are unavailable, but estimation of prediction intervals on the complete estimated continuous total daily and longer-term loads is not currently possible.

In 2011, as part of GLRI, the USGS began a monitoring program to describe concentrations and loads of nutrients and sediment in 30 U.S. Great Lake tributaries. At each site, continuous flow and WQ surrogates (turbidity, temperature, specific conductance, pH, and dissolved oxygen) are measured, and discrete samples are collected and analyzed for selected nutrient and sediment constituents. Herein, we describe this monitoring program (sites and sampling protocols), and then describe how the data are being used to estimate continuous nutrient and sediment loading (and corresponding prediction intervals) from each tributary during 2011–13 using continuous surrogate regression techniques. This monitoring effort provides baseline conditions and information to quantify progress of the various GLRI restoration efforts. The regression equations developed in this study can also be used to estimate continuous loads in the monitored tributaries after 2013 from more recently collected flow and WQ surrogate data as long as the equations are periodically re-evaluated.

Methods

Monitoring program design

Monitoring sites were selected in 30 U.S. tributaries (Table 1; Fig. 1), which were a subset of the 59 tributaries identified in the National