Journal of Hydrology 547 (2017) 575–584

Contents lists available at ScienceDirect

Journal of Hydrology

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

Research papers

Capacity of semi-parametric regression models to predict extreme-event water quality in the Northeastern US

Mark Hagemann, Mi-Hyun Park *

Department of Civil and Environmental Engineering, University of Massachusetts Amherst, United States

ARTICLE INFO

Article history:
Received 8 April 2016
Received in revised form 10 February 2017
Accepted 11 February 2017
Available online 16 February 2017

Keywords:
Extreme events
Model validation
Load estimation
Uncertainty

ABSTRACT

This study assessed the capacity of semi-parametric regression models to predict riverine solute concentrations during extreme high-flow hydrologic events, when such events are absent from the models' calibration data. Using a large dataset from 459 monitoring stations across the US Northeast, the models showed a tendency to overpredict extreme-event concentrations, with increasing bias and variance for increasingly extreme hydrologic conditions. The validation framework in this study effectively compared model performance across disparate hydrologic regimes and constituents, yet can be used to estimate individual model performance under an unobserved extreme-flow condition, regardless of whether any extreme-flow data are available for that model. The validation procedure can further be generalized to explore model performance in an arbitrarily defined extreme condition for a broad range of model types. Despite an overall increase in uncertainty for extreme-event concentration estimates, estimates under extreme hydrologic conditions could be improved by taking into account the observed bias in the aggregated regional database.

1. Introduction

Concentration and mass flux of riverine constituents are two environmental parameters that are of high social and environmental interest, yet difficult to measure with satisfactory time resolution. This has led to widespread modeling efforts that attempt to fill in the resulting gaps in the observed time series, typically using regression models –often referred to as "rating curves"– that estimate log-transformed concentration or mass flux as a function of more easily measured variables including discharge.

The empirical (and explicit, in the case of mass flux) relationships with hydrological variables on which such models rely lead to a natural focus on moments of large variations in flow, i.e. on storm conditions. These periods constitute transport "hot moments" (Vidon et al., 2010) for many riverine constituents, and are responsible for transporting over half of all mass loads in many cases (Raymond and Saiers, 2010). Several recent extreme high-flow events in the US Northeast, including Hurricanes Irene (August 2011) and Sandy (October 2012) have prompted increased scientific attention on the impacts such events have on transport of constituents including nutrients (Yoon and Raymond, 2012), organic matter (Caverly et al., 2013; Dhillon and Inamdar, 2013, 2014), and suspended sediment (Yellen et al., 2014). Several of these studies point to disproportionately large exports during such events, exceeding model predictions. However, these studies focus primarily on quantifying exports from individual storms and do not make a systematic assessment of model performance under such extremes.

A guiding principal for hydrologic modeling was stated in Klemeš (1986) – "Before it is used operationally, a model must demonstrate how well it can perform the kind of task for which it is intended". As modelers increasingly seek to predict impacts of previously unobserved weather and climate conditions (e.g. Carpenter et al., 2015), empirical constituent models may be used to predict water-quality responses to a hypothetical extreme storm event, or to estimate unmeasured conditions during an actual event. Other environmental modeling disciplines have attempted to establish the range of climatic conditions under which their models yield acceptable results (Andréassian et al., 2009; Coron et al., 2012), but to date no systematic assessment has been made of rating-curve models under extreme hydrologic conditions.

This study addresses the question of how well a rating-curve model makes predictions in extreme-flow conditions, given that such conditions are beyond the range seen in its calibration data. Before addressing this question, it is useful to lay out some of the assumptions upon which rating-curve models rest.

* Corresponding author.
E-mail address: mpark@ecs.umass.edu (M.-H. Park).

http://dx.doi.org/10.1016/j.jhydrol.2017.02.017
0022-1694/

- The mean (μ_{lnC}) of the random variable representing log-transformed concentration (lnC) is a function of log-transformed flow and other variables such as season and time. The mean can thus be written conditionally on a set of measurable variables X: $\mu_{lnC|X}$. As predictors in a regression model, these variables explain a portion of the variance in the modeled quantity (i.e. the "response variable"); knowing the value of the predictors reduces the uncertainty in the response. In most cases, the functional relationships are assumed to be linear or quadratic with respect to a transformation of the predictors (e.g. logarithm for flow, harmonic for season; (Cohn et al., 1992)). These assumptions can lead to model bias where they are incorrectly applied (Hirsch, 2014), and other functional forms have been introduced in extensions of the linear model (Autin and Edwards, 2010; Hirsch et al., 2010; Wang et al., 2011).
- The variance (σ^2_{lnC}) and standard deviation (σ_{lnC}) of log-transformed concentration are constant for all times and flow conditions. In practice, some datasets have shown this to be a poor assumption, potentially biasing the estimated parameters of the resulting rating curve. Concentration is typically assumed to follow a log-normal distribution, i.e. $lnC \sim N(\mu_{lnC}, \sigma^2_{lnC})$ (Esmen and Hammad, 1977; Helsel and Hirsch, 2002); this facilitates bias correction of concentration and flux estimates when retransforming from log-space.

One mathematical consequence of these assumptions is that the standard deviation of concentration ($\sigma_{C|X}$) is directly proportional to the conditional mean, implying that larger estimates have larger uncertainty. In the case of log-normality, the proportionality constant grows exponentially with increasing σ^2_{lnC}, the variance of lnC, about its conditional mean: $\sigma_{C|X} = (e^{\sigma^2_{lnC}} - 1)^{\frac{1}{2}} \mu_{C|X}$. As a result, concentration estimates during large hydrologic events are inherently more uncertain than those for less extreme events, whereas estimates of log-transformed concentration have similar precision for all conditions.

The asymmetric distribution of the concentration random variable introduces a bias in retransforming predictions of lnC from log-space. Therefore, while model validity in predicting lnC is a necessary condition for validity in predicting concentration (and by extension, mass load), it is not a sufficient one. Various methods exist to address retransformation bias, some relying on distributional assumptions and others that do not (Cohn et al., 1989). Nonetheless, the performance of a rating-curve model as explicitly defined—as a predictor of lnC—must first be established before considering retransformation bias.

This study therefore sought to evaluate whether semiparametric rating-curve predictions of lnC retain their predictive capacity in extreme high-flow events, relative to their performance in less extreme conditions. As such, metrics of model performance (e.g. bias, goodness-of-fit) are defined herein with respect to log-transformed concentration, rather than the retransformed values of interest. We pay specific attention to the supposition of thresholds beyond which constituent behavior undergoes fundamental changes (Dhillon and Inamdar, 2013, 2014), rendering models inaccurate (Yoon and Raymond, 2012). We further make recommendations about the collection, management, and dissemination of water-quality data in order to improve large-scale data-driven studies.

2. Materials and methods

2.1. Data acquisition

Concentration data for streams in the US Northeast were obtained from the National Water Quality Monitoring Council Water Quality Portal (http://www.waterqualitydata.us/). The initial database query extracted all water-quality data for selected constituents from stream monitoring stations located in the Northeast US between 36°N and 48°N latitude and between 81°W and 66°W longitude (Fig. 1). Daily discharge data for water-quality monitoring sites were obtained from the US Geological Survey (USGS) National Water Information System (NWIS). The data were filtered to include only the datasets with at least 30 concurrent measurements of concentration and discharge for a given station and constituent. A total of 2747 datasets were obtained from 459 monitoring stations, with each dataset representing a unique combination of constituent (nutrients such as nitrogen and phosphorus, organic carbon, and total suspended solids), fraction (dissolved, suspended, or total), and monitoring station (Table 1). In some cases where it was not explicitly provided the "suspended" fraction was calculated from the difference between "total" and "dissolved" fractions, while the fractions of certain dissolved constituents reported as "total" were discarded following USGS recommendations (Rickert, 1992). Each dataset contained between 31 and 3098 concurrent (same day) observations of concentration, daily average flow, and date of measurement.

2.2. Model development

A semiparametric rating-curve model (Wang et al., 2011; Kuhnert et al., 2012) was calibrated to each discharge-concentration dataset using the *gam* function in the *mgcv* R package (Wood, 2011). This model is similar to a traditional rating-curve, in which log-transformed concentration is linearly regressed on log-transformed discharge and other variables representing seasonal and long-term fluctuations. The main difference is that functional relationships may be arbitrarily nonlinear. The model has the form

$$lnC = s_1(lnQ) + s_2(doy) + s_3(time) + \in \quad (1)$$

where lnC is log-transformed concentration; lnQ is log-transformed flow; doy is the numeric day of the year, (1–365 or 1–366); $time$ is the time of observation in days from the mean observation time; and $\in$ is a zero-mean, constant-variance error term. The functions $s_1()$, $s_2()$, and $s_3()$ are nonparametric smooth-functions based on cubic splines (Wood, 2006). Each smooth function includes an additional parameter, k, dictating the maximum degrees of freedom allowed in the function. The k values used in this study restricted lnQ and doy to each have a maximum of 3 degrees of freedom, while an automatically selected default k value was used for time (Wood, 2003). As a comparison, the "7 parameter" Load estimator (LOADEST) model (Cohn et al., 1992) allows 2 degrees of freedom for each lnQ and (harmonically transformed) doy. An advantage of GAMs is that they seldom use the full allowed degrees of freedom; models in this study had a median "effective" degrees of freedom of 8.8, slightly greater than the "7-parameter" LOADEST model's 7 degrees of freedom, with the lnQ term using 2.1 effective degrees of freedom on average. The nonparametric smooth functions behave similarly to (but mathematically differ from) locally weighted regression, such as that employed in Weighted Regression on Time, Discharge, and Season (WRTDS) (Hirsch et al., 2010). Contrary to WRTDS, the smooth functions $s_1()$, $s_2()$, and $s_3()$ are assumed independent of one another. Although other rating-curve models have used covariates other than time, season, and discharge, these three are by far the most commonly used (Hirsch, 2014).

2.3. Differential split-sample test

In order to simulate model performance in a previously unobserved extreme hydrologic condition, a calibration-validation