University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

USGS Staff -- Published Research | US Geological Survey

2014

LARGE BIASES IN REGRESSION-BASED CONSTITUENT FLUX ESTIMATES: CAUSES AND DIAGNOSTIC TOOLS

Robert Hirsh
U.S. Geological Survey, rhirsch@usgs.gov

Follow this and additional works at: http://digitalcommons.unl.edu/usgsstaffpub

Hirsh, Robert, "LARGE BIASES IN REGRESSION-BASED CONSTITUENT FLUX ESTIMATES: CAUSES AND DIAGNOSTIC TOOLS" (2014). *USGS Staff -- Published Research*. 826.
http://digitalcommons.unl.edu/usgsstaffpub/826

This Article is brought to you for free and open access by the US Geological Survey at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in USGS Staff -- Published Research by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

AMERICAN WATER RESOURCES ASSOCIATION

LARGE BIASES IN REGRESSION-BASED CONSTITUENT FLUX ESTIMATES: CAUSES AND DIAGNOSTIC TOOLS[1]

Robert M. Hirsch[2]

ABSTRACT: It has been documented in the literature that, in some cases, widely used regression-based models can produce severely biased estimates of long-term mean river fluxes of various constituents. These models, estimated using sample values of concentration, discharge, and date, are used to compute estimated fluxes for a multiyear period at a daily time step. This study compares results of the LOADEST seven-parameter model, LOADEST five-parameter model, and the Weighted Regressions on Time, Discharge, and Season (WRTDS) model using subsampling of six very large datasets to better understand this bias problem. This analysis considers sample datasets for dissolved nitrate and total phosphorus. The results show that LOADEST-7 and LOADEST-5, although they often produce very nearly unbiased results, can produce highly biased results. This study identifies three conditions that can give rise to these severe biases: (1) lack of fit of the log of concentration *vs.* log discharge relationship, (2) substantial differences in the shape of this relationship across seasons, and (3) severely heteroscedastic residuals. The WRTDS model is more resistant to the bias problem than the LOADEST models but is not immune to them. Understanding the causes of the bias problem is crucial to selecting an appropriate method for flux computations. Diagnostic tools for identifying the potential for bias problems are introduced, and strategies for resolving bias problems are described.

(KEY TERMS: nutrients; transport and fate; statistics; computational methods.)

Hirsch, Robert M., 2014. Large Biases in Regression-Based Constituent Flux Estimates: Causes and Diagnostic Tools. *Journal of the American Water Resources Association* (JAWRA) 1-24. DOI: 10.1111/jawr.12195

INTRODUCTION

A long-standing challenge in the study of water quality is the estimation of the flux of a suspended or dissolved substance in a river, averaged over a time period such as a year or a decade. This is commonly known as the load estimation problem. The problem setting is the following. At a monitoring location on a river there is a set of instantaneous measurements of concentration. They are fairly sparsely measured in time, for example, 12-36 observations per year, measured over some study period of about a decade. They are accompanied by a complete record of daily mean discharge values for the full study period for a location at, or very near, the sample collection site.

A common method for estimating the average flux values for monthly, annual, or multiyear periods is to use multiple regression to estimate a daily flux based on observations of daily discharge, time, season, and in some cases other variables derived from these. These daily flux estimates are then summed to form estimates of average flux over the period of interest. These types of models are generally referred to as "rating curve" or "regression-based" approaches. There is an extensive literature about these models that includes

[1]Paper No. JAWRA-13-0194-P of the *Journal of the American Water Resources Association* (JAWRA). Received September 10, 2013; accepted February 25, 2014. © 2014 American Water Resources Association. This article is a U.S. Government work and is in the public domain in the USA. **Discussions are open until six months from print publication**.

[2]Research Hydrologist, U.S. Geological Survey, 432 National Center, USGS, Reston, Virginia 20192 (E-Mail/Hirsch: rhirsch@usgs.gov).

the following: Dolan *et al.* (1981), Ferguson (1986, 1987), Cohn *et al.* (1989, 1992), Preston *et al.* (1989), Crawford (1991), Robertson and Roerish (1999), Runkel *et al.* (2004), Cohn (2005), Crowder *et al.* (2007), Hirsch *et al.* (2010), Stenback *et al.* (2011), Verma *et al.* (2012), and Richards *et al.* (2012). This study will examine three examples of such regression-based estimates. These are the seven-parameter LOADEST model (L7), the five-parameter LOADEST model (L5), and Weighted Regressions on Time, Discharge, and Season (WRTDS). The LOADEST models are perhaps the most commonly used approaches, and they have been used in many applications in the U.S. Geological Survey (USGS) over the past two decades including the estimation of mean flux values used in Spatially Referenced Regressions on Watershed Attributes (see Smith *et al.*, 1997; Preston *et al.*, 2011). The WRTDS method was introduced more recently (Hirsch *et al.*, 2010), in response to some of the limitations of LOADEST and has been applied in studies of the Mississippi River, Lake Champlain, and Chesapeake Bay watersheds (see Sprague *et al.*, 2011; Medalie *et al.*, 2012; Zhang *et al.*, 2013). There are other variations on the LOADEST model that also address some of its weaknesses. Examples include a variety of models with other explanatory variables (e.g., Hirsch, 1988; Vecchia *et al.*, 2009; Garrett, 2012). All of these include the use of explanatory variables beyond the three types of variables used in LOADEST and WRTDS (time, discharge, and season) but they are not considered in this study. Reliable parameter estimation may be difficult for such models if sample sizes are small. However, such models may be useful as potential solutions to the bias problem. The diagnostic approaches proposed in this study for the L5, L7, and WRTDS models would all be applicable to these more complex models.

All three of these models share the same motivation: use some version of multiple regression to estimate the concentration (and hence flux) on unsampled days. In each case, the regression assumes that the log of concentration is the sum of four components drivers: discharge, season, long-term trend, and random unexplained variation. These models should not be thought of as a simple causative model. Rather, they are used because in many situations they can provide efficient and unbiased estimates of concentrations on unsampled days. If this formulation based on discharge, season, and time does not remove substantial amounts of variance from the data, then other methods, such as interpolation or ratio estimators (see Richards and Holloway (1987) for a description of the Beale Ratio Estimator) may be better estimators. Purely deterministic models are also available for estimating mean flux. These models are not considered here. The datasets considered in this study all have characteristics that suggest that this type of multiple regression procedure is potentially useful for estimating fluxes on all days, and hence estimating long-term mean fluxes.

There have been several articles published in the last three years that explore the bias of these types of flux estimates. These include Stenback *et al.* (2011), Garrett (2012), Moyer *et al.* (2012), and Richards *et al.* (2012). A 2013 update to the USGS LOADEST code also discusses this issue and provides diagnostics for the bias problems (at http://water.usgs.gov/software/loadest/doc/loadest_update.pdf). An important point made in all of these studies is that there are cases in which application of some of these regression-based approaches can produce long-term average flux estimates that are biased by many tens of percent and either positive or negative in sign. These studies also show that there are many cases in which one or more of these models provide estimates that are virtually unbiased. The studies that identify the potential for severe bias have only brief discussions of potential causes of these problems. Within these discussions, heteroscedastic residuals are mentioned in the first three, lack of fit is mentioned in the last two, and failure to properly capture the seasonal pattern is mentioned only in the first of these.

The goal of this study was to advance the understanding of the problem of large biases in these three regression-based flux estimates. This includes discussion both of the statistical issues as well as some of the hydrologic processes that may give rise to these problems and also suggest some methods that practitioners can use to help identify cases in which these problems may exist. This will be accomplished through the analysis of a limited collection of datasets, which are sufficiently rich in samples such that the true fluxes for decade-long period can be approximated directly from the data. Small samples are then selected from these datasets to evaluate the bias in estimates that would have been computed by using each of these regression-based models. The analysis will show that these bias problems arise due to severe violations of the set of assumptions that are the basis of the L5 or L7 models. It will also show that these problems may be reduced or eliminated under the much less restrictive assumptions of the WRTDS model. The study describes some tools for diagnosing the problem and suggests approaches to reducing the risk of producing highly biased results.

This study only considers datasets of 120 observations or more, which are representative of the full distribution of discharges for the site, and have no censoring. All three models do include appropriate computational schemes that allow for the analysis of censored data, but to limit the complexity of the analysis, censored cases were not considered here.

DESCRIPTION OF THE FLUX ESTIMATION MODELS

The LOADEST model forms the basis for all three models considered here (L5, L7, and WRTDS). The LOADEST computer package (Runkel *et al.*, 2004) is one standard implementation of both the L5 and L7 models considered here. The following section will describe the L7 model. Based on that, the L5 and WRTDS will be described as variations on the basic concept in L7. All three of these methods share the same premise that there is predictive power gained by attempting to model the behavior of concentration as a function of the three explanatory variables: discharge, time, and season. If the predictive power is very poor it may be that other estimation approaches such as linear interpolation in time or ratio estimators should be considered. Comparison of these regression-based approaches with these alternatives is beyond the scope of this study.

The L7 Model

The L7 model is based on Cohn *et al.* (1989, 1992) and Cohn (2005) and the relevant software is documented by Runkel *et al.* (2004). The approach is based on "...a particular linear model [that] satisfactorily describes much of the variability of constituent concentrations" (Cohn *et al.*, 1992). The model has seven fitted parameters and can be written in the following form:

$$\ln(c) = \beta_1 + \beta_2\left[\ln(Q) - \overline{(\ln(Q))}\right] + \beta_3\left[\ln(Q) - \overline{(\ln(Q))}\right]^2 + \beta_4(T - \bar{T}) + \beta_5(T - \bar{T})^2 + \beta_6\sin(2\pi T) + \beta_7\cos(2\pi T) + \varepsilon \quad (1)$$

where ln() is the natural logarithmic function; c is the concentration; Q is the discharge; $\overline{(\ln(Q))}$ is the mean of the natural log discharge values on the sampled days; T is time in decimal years; $\bar{T}$ is the mean value of time in decimal years for the sampled days; ε is the error, assumed to be normally distributed with constant variance of σ_ε^2; β_1, β_2.....β_7 are the coefficients, estimated from the sample data.

The LOADEST program allows for several specific estimation methods, including the use of fewer, or more, explanatory variables that are shown in this equation. The L7 model uses multiple linear regression to fit the parameters of Equation (1) to the available data and then applies the fitted model to estimate values of ln(c) for all of the days in the complete period of study, using discharge and time as the explanatory variables. The resulting estimates of ln(c) are then back-transformed (by exponentiation) to concentration units and these are multiplied by a bias correction factor (BCF). The BCF is intended to compensate for the fact that the expected value of concentration is not simply the back-transformed value of the expected value of ln(c). The method for computing the BCF is described by Cohn (2005). These daily estimates of concentration are then multiplied by the daily discharge values and a unit conversion factor to obtain an estimate of daily flux. These daily estimates can then be summed to form estimates of monthly, annual, or long-term average values of flux.

Cohn (2005) demonstrated that Equation (1) provides unbiased estimates of flux if all of the model assumptions are met. The practical issue, however, is to determine how sensitive the results are to departures from the idealized behavior presented in Equation (1). Some of the possible types of departures that can be found in long-term datasets include the following: relationships between ln(c) and ln(Q) that do not follow the quadratic functional form, trends that do not follow a quadratic functional form, seasonal patterns that do not follow a sinusoidal pattern, seasonal patterns that change over the range of time or discharge (changing their amplitude or phase shift or overall shape), interactions between terms of the model (for example, trends that are different for high discharges than for low discharges, or are different in different seasons), nonnormal distribution of the error term, or an error variance that changes as a function of one or more of the explanatory variables. Hirsch *et al.* (2010) provide examples of several of these problems.

Cohn *et al.* (1992) applied the L7 model to 24 nutrient datasets of nine-year duration from four rivers in the Chesapeake Bay watershed (drainage areas that range from 182 to 43,600 km^2). Cohn *et al.* (1992) state the following conclusions about the applicability of this model to these datasets: The model described by Equation (1) "...provided a useful and reasonably accurate description of nutrient concentrations *in the streams examined here*. However, statistically significant, though not substantial, lack of fit was observed in all cases. Load estimates assuming the validity of [this model] appears to be fairly insensitive to modest amounts of model misspecification or nonnormality of residual errors.... In summary, the...load estimator...was found to provide satisfactory estimates both of nutrient loads and of the uncertainties to total load estimates" (emphasis added).

Although Cohn *et al.* (1992) formally tested the applicability of the model using 24 datasets, no claim was made in that study or in the LOADEST documentation as to the universal applicability of the model to all concentration and discharge datasets. In particular, the claim that L7 is unbiased must be understood to mean that it is unbiased in the case where the actual river system conforms to the assumptions presented by

Cohn *et al.* (1992). Guo *et al.* (2002), Stenback *et al.* (2011), Garrett (2012), and Moyer *et al.* (2012) all present examples of cases in which the data demonstrate substantial departures from one or more of those assumptions, and as a consequence of these departures from the model assumptions, they result in flux estimates that are clearly biased. In summary, the literature correctly indicates that this model is unbiased when the model assumptions are, at least, approximately valid but that situations can arise in which severe departures from the assumptions can lead to severe biases. One of the chief aims of this study was to provide some guidance to help the hydrologist identify departures that are sufficiently severe to cause one of the methods to result in substantially biased results.

The L5 Model

The L5 model is the same as L7 except that two of the explanatory variables are eliminated from consideration, making it a five-parameter model instead of a seven-parameter model by deleting the two quadratic terms. The model can be written in the following form:

$$\ln(c) = \beta_1 + \beta_2(\ln(Q)) + \beta_3(T) + \beta_4 \sin(2\pi T) + \beta_5 \cos(2\pi T) + \varepsilon \quad (2)$$

Like the L7 model, the L5 model includes the use of a BCF to account for the retransformation bias. It is computed by the same method that is used in the L7 model. The L5 approach is sometimes favored because of a concern that the curvature derived from the quadratic terms may result in very extreme estimates near or beyond the limits of the sampled values of $\ln(Q)$ or T in the dataset. That is a reasonable cause for concern, particularly when the daily discharge dataset extends to discharge values that are substantially higher than those in the water quality sample dataset. However, balanced against this concern, one must also consider the case where there is demonstrable curvature in the relationship between $\ln(Q)$ and $\ln(c)$, in which case the lack of the quadratic term in $\ln(Q)$ could result in severe underestimation or overestimation at high discharges, depending on the direction of the curvature. The issue of flux bias for the L5 model has been discussed by Richards *et al.* (2012).

The WRTDS Model

The method, Weighted Regressions on Time, Discharge, and Season (WRTDS) was introduced by Hirsch *et al.* (2010) and have been applied in several recent water quality studies (Medalie, *et al.* 2012; Sprague *et al.*, 2011; Hirsch, 2012; Moyer *et al.*, 2012; Zhang *et al.*, 2013). It also uses the same predictive equation as the L5 model (Equation 2). The estimation method differs from either L7 or L5 in two very fundamental respects. The first is that the coefficients are not fixed, but vary in a gradual manner throughout the Q,T space. This is accomplished by applying weighted regression for the estimation of $\ln(c)$ where the weights are established as a function of the "distance" between the estimation point (defined by Q and T) and the sample points. The measure of "distance" is defined in three dimensions: $\log(Q)$, T, and season (proximity to the same time of year). Thus, the relationship of $\ln(c)$ is considered to be locally linear in $\ln(Q)$, T, $\sin(2\pi T)$, and $\cos(2\pi T)$ but more generally the estimates of $\ln(c)$ are defined by a continuous smooth surface in Q,T space. The weights used in the weighted regression are the product of the three individual weights (based on discharge, time, and season). The use of weighted regression is motivated by the idea that estimates for a particular year, season, and discharge should not be highly influenced by observations that were collected under very different conditions. For example, data collected in July should have little or no influence on estimates for January or data collected at a discharge of 10 m^3/s should have little or no influence on estimates for conditions at 1,000 m^3/s. The second way that it differs is in how it determines the BCF. In L5 and L7 the assumption is made that the errors are normal and homoscedastic, and as such the BCF is approximately equal to $\exp(SE^2/2)$ where SE is the standard error of the residuals of Equation (1). In WRTDS the assumption is made that the errors are normal but are heteroscedastic, and as such standard error is not considered to be a constant but is assumed to vary gradually over the Q,T space. The standard error, for any combination of Q and T, is estimated by the same weighted multiple regression method that is described above. Experience with many nutrient datasets, particularly for nitrate, exhibit much higher variability of $\ln(c)$ under summer low-flow conditions than for other seasons or flow conditions. For a fuller discussion of WRTDS see Hirsch *et al.* (2010) and for the extension to the censored case see Moyer *et al.* (2012). Like the L5 and L7 models the actual estimation is done using a weighted Tobit model (Tobin, 1958) to accommodate censored values, but when there are no censored values, this model becomes virtually equivalent to weighted multiple linear regression.

DESIGN OF THE RESAMPLING EXPERIMENT

During the research leading up to the writing of this study many datasets were examined by each of

the models and the findings showed that there are many cases in which all or most of the models performed very well, exhibited little or no bias, and often result in nearly equivalent results. But, the problem of substantial bias is common enough that it warrants having the hydrologist take care to determine if it might be a problem for the dataset and model at hand. To provide more insight on this problem an experiment was devised in which the datasets were rich enough that a mean flux could be computed rather accurately from the data, without the need for any regression-based estimates, and then these datasets could be subsampled and the regression models estimated and used to compute mean flux which could then be compared to the true value. The experiment considers six specific cases. Each case is defined by a dataset of dissolved nitrate plus nitrite (denoted here for simplicity as NO_3) or total phosphorus (TP) at a specific sampling location. Exploration (R.M. Hirsch, unpublished) of a wider range of analytes suggests that these two are among the most problematic and their specific problems arise for reasons related to different kinds of processes. Suspended sediment is likely to be subject to many of the same issues as TP and is known to be difficult to estimate very accurately because of its highly nonlinear relationship with discharge. Other constituents of importance, which could present serious bias issues, are those that commonly have a high degree of censoring (these include orthophosphorus, pesticides, metals, and organic compounds). These were not considered here. The datasets used here primarily come from the long-term monitoring program of Heidelberg University (in Tiffin, Ohio) and one comes from the Des Moines Iowa Water Works. The diligence of these two organizations in collecting, quality assuring, storing, and making these rich datasets publicly available is gratefully acknowledged.

The resampling experiment has several goals. (1) Determine if there is one model that is consistently the least biased of the three; (2) Determine if there is one model that is consistently the most biased of the three; (3) Determine if there is one model that is relatively robust (that is, either best or close to the best from a bias perspective); (4) Determine if the bias problem is strongly related to sample size; (5) Determine if the bias problem can be reduced by use of a stratified sampling strategy; (6) Create a large set of examples to test the efficacy of a practical diagnostic statistic (introduced below); and (7) To provide some context for selecting a few example cases that help illustrate some of the causes of the bias and the utility of a set of diagnostic graphics. The resampling experiment uses only 10 repetitions for each case and sampling frequency. The small number of repetitions is consistent with the goal of the study, which is to address large differences in bias across the methods, as opposed to hypothesis testing about these differences. The focus is on practical significance and not on statistical significance.

The design of the resampling experiment is this:

1. For each case use the full dataset to determine the "true" flux on a large number of the days over the period of record. In most cases this was about 90% of the days although in one case (the Iowa case) it is only 45% of the days. These true fluxes are summarized into annual mean values and these are then summarized into period-of-record mean values. These annual mean values are the mean for all of the sampled days and not a mean for all 365 days of the year. It is recognized that this "true flux" is not a perfect estimate of the daily flux (given issues of measurement and sampling accuracy as well as variations in concentration and discharge that take place at the subdaily time scale). But, they provide a high standard of accuracy, to which the inherently less accurate regression-based estimates can be compared. The specific algorithm for computing the true flux for the day is described in Appendix S1.
2. A random sample is drawn, without replacement, from the full set of samples. These have average sampling rates that are roughly weekly, biweekly, or monthly, although they are random rather than being taken at a standard time step. A stratified random sampling scheme is also applied and those results are shown in Appendix S2. The stratified random sampling scheme was designed as a rough approximation to the approach often used by the USGS and some other agencies, whereby high discharge days have a higher probability of being sampled than days with low to moderate discharge.
3. Each random sample is used to estimate each of the models (L5, L7, and WRTDS) and the resulting model is used to estimate the flux on every day of the period of record. These estimates all incorporate the use of the appropriate BCF to compensate for the retransformation bias.
4. For those days for which a true flux is known, the daily estimates are summed to form estimates of the annual mean flux (again, this is a mean that excludes the days that were not sampled in the full record). From these estimates of annual flux a long-term mean annual flux is estimated.
5. For each iteration (defined by a particular random sample from a given dataset, using a given

model) an error is computed, which is the estimated long-term mean minus the true long-term mean. In both cases these mean values are computed for those days for which the true flux is known. A standardized form of this error is computed by dividing by the true long-term mean. Finally, this value is multiplied by 100, so that it can be expressed as error in percent. This standardized error is denoted as E_m. Steps 4 and 5 are described in more detail in Appendix S1.

6. For each iteration the "Flux Bias Statistic" (B_m) is computed from the results. This diagnostic statistic is functionally equivalent to the "partial load ratio" used by Stenback *et al.* (2011). It is a dimensionless representation of the difference between the sum of the estimated fluxes on all sampled days (P_m) and the sum of the true fluxes on all sampled days (O). Ideally, one would hope that it is a good indicator of the true flux (E_m) which can only be known when data from all days are available. B_m is defined as follows.

$$B_m = (P_m - O)/P_m \tag{3}$$

where

$$O = \sum_{i=1}^{n} L_i = \sum_{i=1}^{n} k \cdot c_i \cdot Q_i \tag{4}$$

$$P_m = \sum_{i=1}^{n} \hat{L}_i = \sum_{i=1}^{n} k \cdot \hat{c}_i \cdot Q_i \tag{5}$$

where L_i is the observed load on the ith sampled day; $\hat{L}_i$ is the estimated load on the ith sampled day; k is a units conversion factor (if concentration is in mg/l and discharge is in m^3/s and load has units of kg/day, then $k = 86.4$); c_i is the measured concentration on the ith sampled day; $\hat{c}_i$ is the estimated concentration on the ith sampled day; Q_i is the discharge on the ith sampled day; and n is the number of sampled days.

A value of B_m near zero suggests that the model is nearly unbiased. Positive value suggests a positive bias, and a negative value suggests a negative bias.

The resampling experiment was applied to six cases, shown in Table 1. The datasets were selected to represent a range of behaviors that have been observed in examinations of a substantially larger group of datasets (R.M. Hirsch, unpublished). They range from one that exhibits virtually no bias problem regardless of model, to ones that have very severe biases with one or more of the models. Severe bias problems are defined here as average biases of greater than 10% in absolute value. Collectively, they are not intended to be representative of typical datasets, but they should not be construed as highly unusual. The bias problems seen in the application of one or more of these models to these datasets are found with sufficient frequency across a range of datasets that the problems found here must be considered common and not just an oddity that happens in very rare cases. There is no clear method for determining how common these problems are, because of the small number of sites with datasets that are sufficiently rich to allow for such an analysis.

Table 1 is a list of the cases in the resampling experiment. Data for the first five come from Heidelberg University. The source for the last one is the Des Moines Water Works.

Figure 1 shows scatter plots of ln(c) *vs.* ln(Q) for random samples from each of these six cases. In each case, the size of the random sample is approximately 260 observations (a range of 254-268). The full dataset, on the order of 4,000 observations, is not shown because the density of data points actually obscures the key features of the data. These random samples of about 260 observations are much more like what the hydrologist would have available in typical applications.

TABLE 1. List of the Cases in the Resampling Experiment. Data for first five come from Heidelberg University. The source for the last one is the Des Moines Water Works.

Case Abbreviations	Full Name of Site	USGS Station Number	Drainage Area in km^2	Period of Record Used, Water Years	Percent Agricultural Land	Percent Urban Land
Maumee TP	Maumee River at Waterville, Ohio	04193500	16,400	1987-1996	90	1
Honey TP	Honey Creek at Melmore, Ohio	04197100	386	1989-1998	86	1
Cuya NO_3	Cuyahoga River at Independence, Ohio	04208000	1,831	1993-2002	30	10
Honey NO_3	Honey Creek at Melmore, Ohio	04197100	386	1978-1987	86	1
Musk NO_3	Muskingum River at McConnelsville, Ohio	03150000	19,223	2002-2011	52	2
Rac NO_3	Raccoon River at Fleur Drive at Des Moines, Iowa	05484900	9,389	1999-2008	92	3

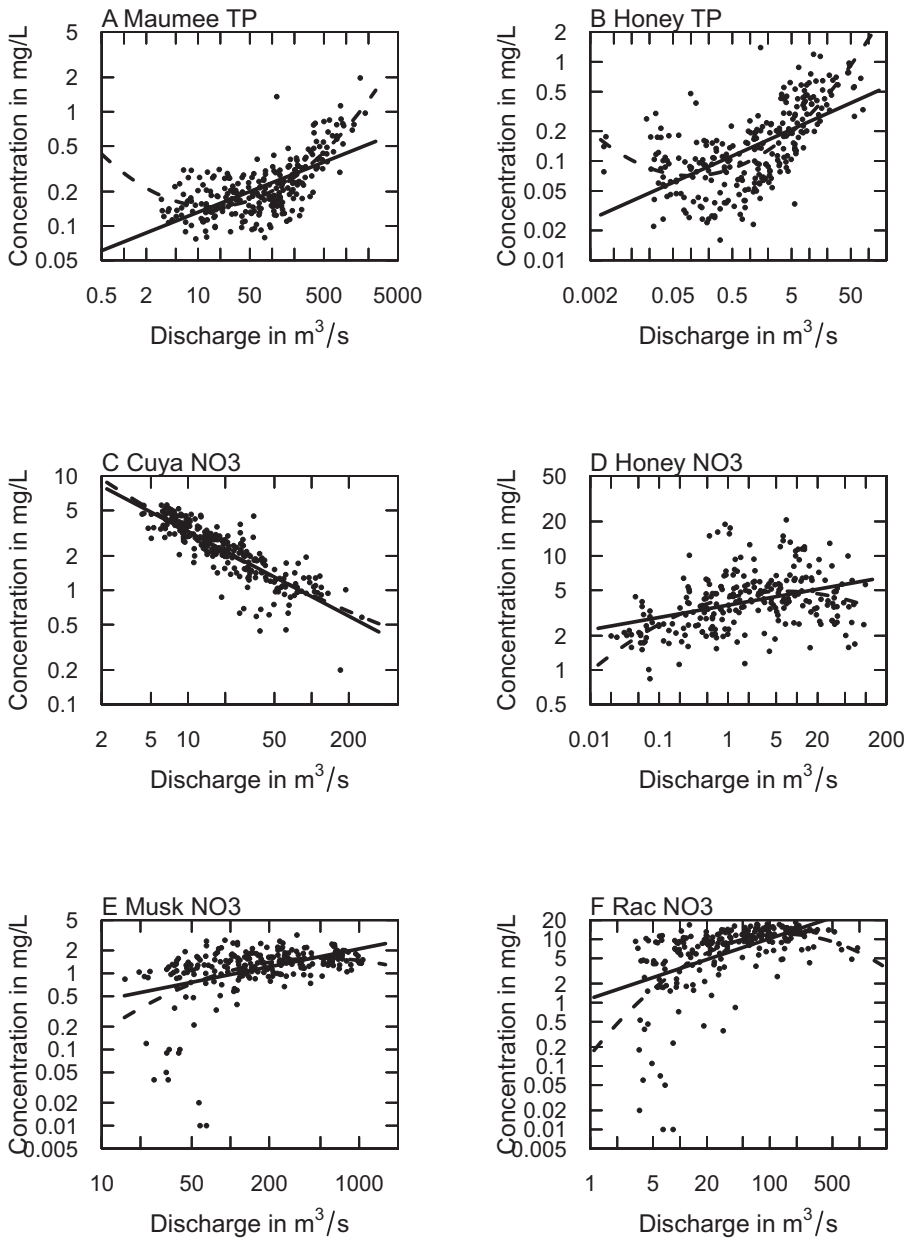

FIGURE 1. Scatter plots of Concentration *vs.* Discharge for Randomly Selected Data from Each of the Six Datasets. Each panel also shows a linear regression fit to the data (solid line) and a linear regression with a quadratic term (dashed line). (A) Maumee TP, (B) Honey TP, (C) Cuya NO_3, (D) Honey NO_3, (E) Musk NO_3, (F) Rac NO_3.

The scatter plots in Figure 1 also show regression fits of a linear model of $\ln(c)$ as a function of $\ln(Q)$ and $\ln(c)$ as a quadratic function of $\ln(Q)$. These can be thought of as simplified versions of the L5 and L7 models, respectively. The simplifications are that they leave out the trend terms and seasonal terms. These plots help elucidate some, but not all, of the common types of problems that give rise to severe biases. A more exhaustive graphical approach is applied in the diagnostic graphics presented in Figures 7-10.

RESULTS OF MONTE CARLO RESAMPLING EXPERIMENT

The resampling experiment was used to evaluate the central tendency and variability of E_m (error in percent) for the six cases listed in Table 1 and illustrated in Figure 1. Figure 2 shows results for the six cases, all three models, and three sample sizes (120, 240, and 480 samples) taken over a 10-year period in all cases. The boxplots each show the E_m values for

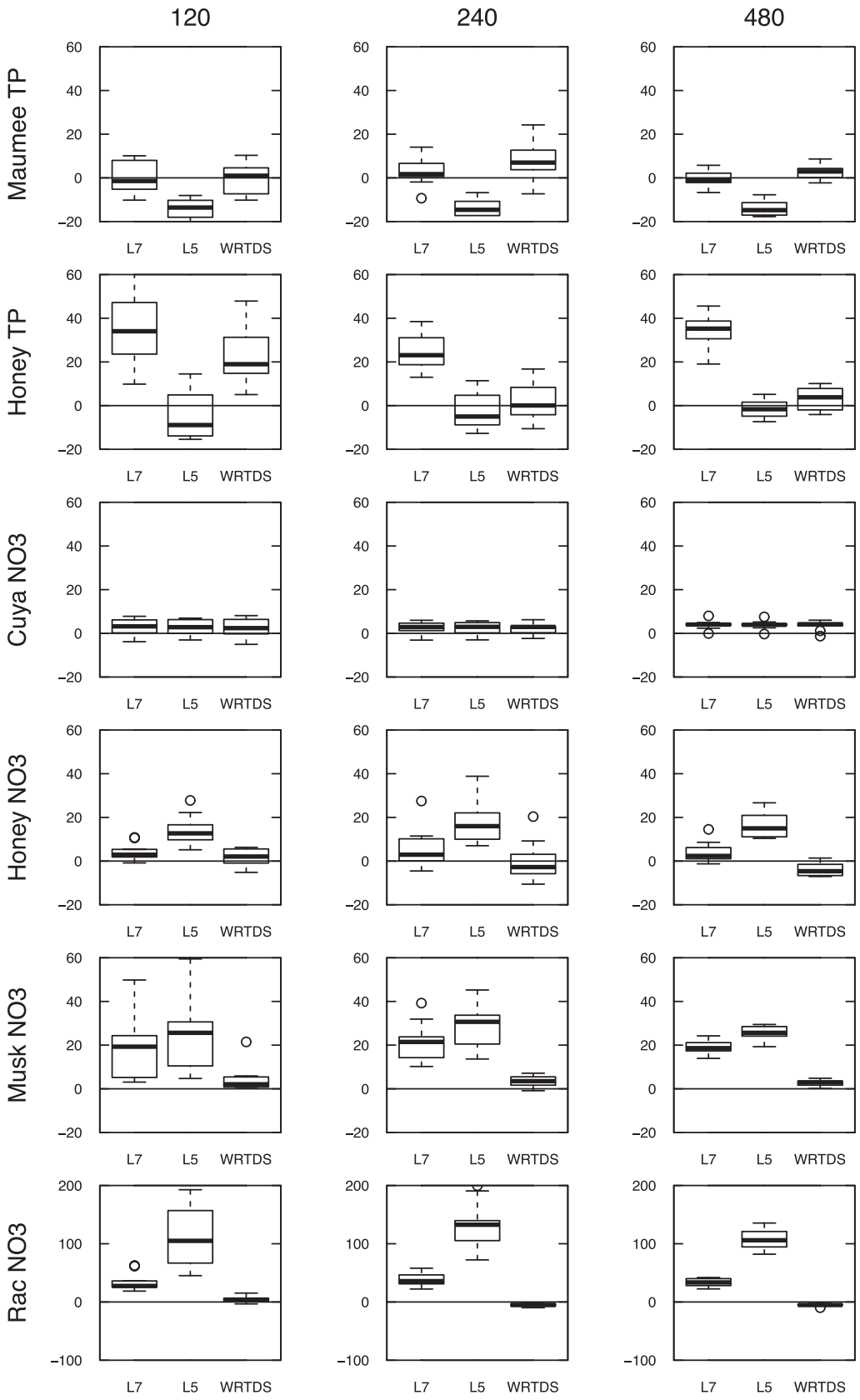

FIGURE 2. Boxplots of the Error in Percent for Estimates of Average Flux, for Six Datasets Selected Using Random Sampling, at Sample Sizes of 120, 240, and 480 Samples, for Three Models (L7, L5, and WRTDS). Each box represents the results of 10 subsamples of the full dataset. Note the scale differences.

10 iterations with a given model, dataset, and sample size. For the three boxplots shown in each panel of Figure 2 the same set of 10 subsamples is used. Thus, the differences among the three boxplots are entirely due to model differences and not to sampling variability. The vertical scaling of all of the panels in the top five rows of the figure is identical. The bottom row of panels represents the Rac NO_3 dataset and it has errors that are much larger than the others. Thus, a different scale is used for it. For an unbiased method the boxplots can be expected to be centered around a value of zero.

Figure 2 speaks to the first four goals of the Monte Carlo experiment. In particular: (1) No one model is consistently the least biased of the three. (2) No one model is consistently the most biased. (3) One model, WRTDS, is relatively robust in the sense that over all 18 comparisons (six cases by three sampling frequencies) in almost all comparisons it is the least biased or virtually indistinguishable from the least biased. The other two models both had instances where the absolute bias was very large compared to the best model. The only exception is for Honey TP at a sample size of 120 observations. The relatively poor performance of WRTDS is likely due to the fact that at the highest discharges there are only a few observations that demonstrate evidence for supply limitation of TP. As a consequence, the WRTDS model does not incorporate the slight downwards slope at high discharges. This behavior is better represented when more data are available (240 or 480 observations). In this one case, the bias for L7 was much more severe than WRTDS, and the L5 was of the opposite sign but was smaller in absolute value. (4) In all cases but Honey TP the bias is largely unrelated to sample size. The precision (depicted by the height of the boxes) almost always decreases with increasing sample size but the bias is rather insensitive to sample size. The fifth goal of the Monte Carlo experiment was to explore if the bias could be reduced by using a sampling strategy that allocates more sampling effort to the highest 20% of discharges. The results from this experiment are shown in Figure S2-1 of the Supporting Information. What it shows is that this kind of stratified sampling does have the effect of reducing the absolute magnitude of some of the most severe biases, but it does not eliminate the bias problem. For example, for Honey TP, with a sample size of 480 and random sampling, the error of the L7 model estimates is about +30 to +40% in most of the repetitions, but with the stratified random sampling the error declines to about +18 to +25%. Overall, application of extra sampling to the higher discharges improves the accuracy of flux estimates but is not a solution to the bias problem. Using a model appropriate to the dataset is fundamentally what is needed to avoid the bias problem.

Summarizing these first five points, we can say that the WRTDS model, while it is more resistant to the bias problem, is not immune to it. This suggests that users need to be mindful of the potential for bias regardless of which model is used. These results also suggest that overall the WRTDS has great advantages in terms of robustness. It never appears to be substantially worse, from a bias perspective, than either of the other two models and there are many situations where it is substantially better.

The next objective of the Monte Carlo experiment was to determine if the flux bias statistic B_m, which is something that the hydrologist can compute from the samples available and the fitted model, is a good predictor of the true bias (E_m), which is not available in practice. To consider this question Figure 3 presents a scatter plot of E_m (true bias) as a function of B_m (the statistic calculated from the sample only) and it uses all of the results from all cases, models, iterations, sample sizes, and sampling strategies (a total

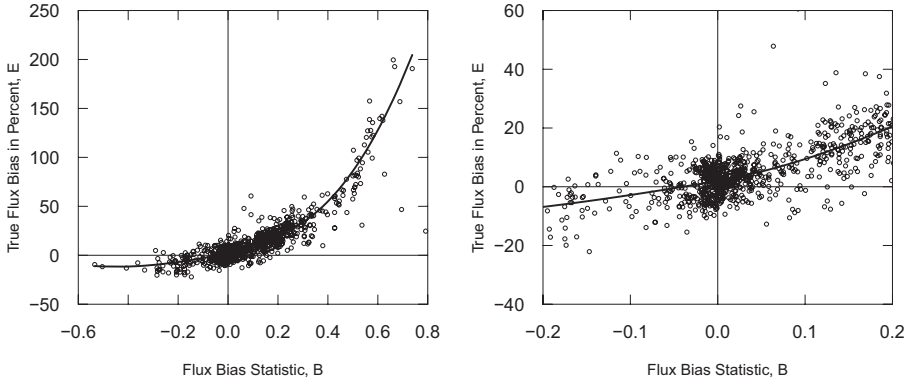

FIGURE 3. Relationship between the True Error in Percent (E_m) and the B_m Statistic Based on the Sample Data, for All 1,080 Cases Considered in the Resampling Experiment. The left panel shows the entire range of B_m values, and right panel provides more detail on B_m values between −0.2 and +0.2. The solid line in the figure is a loess smooth of the scatter plot.

of 1,080 data points). No substantial differences were seen in this relationship when categorized by model, by sample size, or by sampling scheme (random or stratified random). Consequently, the figure shows all 1,080 cases together.

The results seen in Figure 3 show that the B_m is a good general indicator of the potential for bias. It shows that to limit the risk of having a bias of greater than plus or minus 10%, an acceptable range of B_m values may be (−0.1, +0.1). But, there needs to be a recognition that even with a B_m value very close to zero, there is a nontrivial chance of having a flux bias that is greater than 10% in absolute value. Certainly for B_m values that are relatively far from zero (say greater than 0.2 in absolute value) it is fairly certain that the flux results will be rather biased and the sign of the bias is likely to be the same as the sign of B_m. The relationship between B_m and E_m is highly nonlinear. For B_m values near 0.6, for example, true biases are in the range of about 100-125%, but for B_m values near 0.2 the true bias is likely to be about 5-25%. One additional note of caution is that none of the cases examined here were ones when high discharge days were underrepresented (or absent) from the sample dataset. The B_m statistic is unlikely to provide even a rough indicator of bias when high discharge days are severely underrepresented in the sample. The lack of empirical data about concentrations at the high end of the discharge distribution makes it essentially impossible to draw conclusions about the bias of any possible model. In conclusion, the B_m statistic can be viewed as a useful indicator of the potential for bias, but is highly imprecise. It should not be considered as a basis for assigning a correction factor to flux estimates as a means of resolving the bias problem. Rather, it is one indicator that other steps may be necessary to mitigate a potential bias problem.

Note that in the computation of B_m and in the graphical methods described below, the estimated values of log(c), or concentration or flux for sampled days are based on a "leave-one-out cross validation" approach. That approach is used here because it is a somewhat more realistic indicator of actual error because each estimate is made without the knowledge of the true value. Appendix S3 discusses the use of the leave-one-out cross validation approach to this problem. Using this approach has a small impact on conclusions for datasets of the size considered here (120 or more observations) but may be more important when applied to smaller datasets. It certainly adds complexity to the computation, but it has been built into the EGRET software that implements the WRTDS method (https://github.com/USGS-R/EGRET/wiki).

DISCUSSION — COMMON CAUSES OF SEVERE BIAS PROBLEMS

Figure 1 provides some insight into how well these datasets conform to the set of assumptions used by each of these models. In particular, it speaks to the shape of the ln(c) *vs.* ln(Q) relationship and to the homoscedasticity assumption. However, all of the models fit their coefficients simultaneously, so any given scatter plot can only provide a partial view of some of the reasons why any of the three models may fail to provide a reasonable statistical representation of each of these datasets. After extensive exploration of these and other water quality datasets (R.M. Hirsch, unpublished) using many diagnostic tools, it appears that there are three dominant types of behaviors that cause serious bias problems. These three problems are: (1) lack of fit of the ln(c) *vs.* ln(Q) relationship, (2) substantial differences is the shape of the ln(c) *vs.* ln(Q) relationship across different seasons of the year, and (3) major departures from homoscedasticity.

Although not encountered in this study, another possible cause of severe biases is that the shape of the ln(c) *vs.* ln(Q) relationship changes substantially over the period of record. This is likely to arise if there has been a major change in the pollutant source characteristics (for example, a major upgrade or elimination of a point source) or if the record is very long and encompasses large changes in land use practices or long-term changes in concentrations of the pollutant in groundwater such that they cause a change in base-flow concentrations. Examples of this type of situation are described in Hirsch *et al.* (2010), specifically the case of the Patuxent River in Maryland which experienced implementation of advanced waste treatment for phosphorus removal, and Alameddine *et al.* (2011), specifically the case of the change in total nitrogen concentrations as a result of Total Maximum Daily Load (TMDL) implementation in the Neuse River, North Carolina.

The following three sections elaborate on the reasons that each of the three causes described above result in serious biases. For purposes of discussion, the problems are treated here in isolation, but in reality they often arise together.

Lack of Fit of the ln(c) vs. ln(Q) Relationship

The L5 model is constrained to approximate this relationship as linear, although it allows the intercept term to vary over time and over seasons. Of these six cases, only one, Cuya NO_3 (Figure 1C),

appears to have a form that could be adequately modeled as linear. The consequence of the poor fit in the other five cases is that the model is likely to underestimate concentrations in some discharge range and overestimate in others. Because the interest in this study is with estimation of flux, the errors at the higher discharges are the most important. Visual inspection of Figure 1 alone would suggest the following tendencies for flux bias in the application of the L5 model: severe underestimation in the Maumee TP and Honey TP (because the relationship steepens at high discharges); no serious problem with Cuya NO_3 (because it is approximately linear); and severe overestimation in the other three NO_3 cases (because the relationship flattens out or even declines slightly at high discharges).

The L7 model constrains the ln(c) *vs*. ln(Q) relationship to be quadratic. In the case of Maumee TP (Figure 1A), although the relationship might best be described as the combination of two linear segments, it is possible that a quadratic equation could represent this reasonably well. The lower segment, which has a slope of virtually zero, suggests that this represents a background level of TP probably consisting of dissolved phosphorus as well as phosphorus attached to very fine particles. The upper segment covers the discharge range at which significant sediment transport takes place and the TP concentrations rise steeply here as discharge increases. In addition to this process perspective it is also statistically problematic that in the fitting process attempts to make estimates for the highest discharge values (for example, those greater than about 500 m^3/s) will be influenced by observations for which the discharge is one or two orders of magnitude lower. Because there is no theoretical basis for the quadratic form of the model, it is difficult to justify allowing these low Q data points to exert strong influence on the fit at high Q. The quadratic is a pragmatic choice, being a simple functional form that allows for curvature with using only two parameters. The WRTDS model which uses a smoothing approach rather than using a single functional form avoids this problem of the estimates at high discharge being influenced by observations taken at very low discharge. In the case of Honey TP (Figure 1B), overall the dataset appears reasonably true to the quadratic form, although the 10 values observed at the highest Q values suggest the possibility of a departure from the quadratic form. In particular, it suggests that above about 40 m^3/s there may be a depletion of available phosphorus, resulting in a decline in concentration with increasing discharge. A quadratic allows for no more than one inflection point (changes in sign of the first derivative). This is one example where the ln(c) *vs*. ln(Q) may be better characterized by a function with two or more inflection points. However, the amount of data that can help characterize the transition point and slope of the curve in this highest discharge range is very limited. Cuya NO_3 (Figure 1C) is well modeled by the quadratic form of the L7 model even though the quadratic term adds very little to the quality of the fit. The quadratic model does not appear to be a good fit for the Honey NO_3 dataset (Figure 1D), but the departure seems most severe at the lowest discharges. From this standpoint the L7 model might be expected to work reasonably well for this dataset and Figure 2 shows that it does. The final two datasets (Musk NO_3 and Rac NO_3) both have characteristics that make them be poorly modeled by the quadratic form. Both suggest only one inflection point, but they also appear to have ln(c) values that rise steeply with ln(Q) up to some threshold, then appear to have near zero slope over a very broad range of ln(Q) values, and finally show a slight indication of a decline at the very highest ln(Q) values. The failure to properly model this decrease (dilution effect) at high flow is a common cause for substantial positive bias in L7 estimates of NO_3.

The WRTDS model, because it has no preestablished functional form, can generally represent the patterns seen in most of these datasets. The greatest challenge for the WRTDS model in terms of fitting these patterns appears in two cases where a very small number of observations, at very high Q, suggest a negative slope of the ln(c) *vs*. ln(Q) relationship (Honey TP and Rac NO_3). The WRTDS model tends to be relatively insensitive to a change in slope that is apparent in only a few data points. However, this shortcoming of poor fits at the extremes of Q values is generally less severe than what happens when the L5 or L7 models are used.

Seasonal Differences in the ln(c) vs. ln(Q) Relationship

The functional forms of the L5 and L7 model both include an assumption that the ln(c) *vs*. ln(Q) relationship is exactly the same in all seasons except for the intercept term which varies cyclically with a period of one year. In L5 and L7, the slope of the ln(c) *vs*. ln(Q) relationship, for any given value of Q, is required to be constant over all times of year. The WRTDS model is not similarly constrained. The following example illustrates how strongly data can depart from this assumption of the L5 and L7 models. Figure 4 shows the Musk NO_3 dataset (shown in Figure 1E), but here the data are subdivided into a cold season (November-April) and a warm season (May-October). This pattern, which is quite typical of many nitrate datasets, shows cold season concentrations

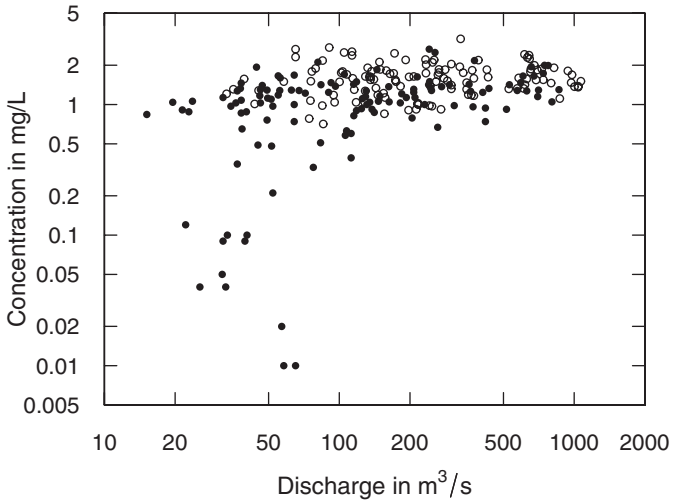

FIGURE 4. Muskingum River, at McConnelsville, Ohio, NO_3 Concentration *vs.* Discharge with Cold Season (November-April) Observations Shown as Open Circles, and Warm Season (May-October) Shown as Closed Circles.

that are almost independent of discharge. Concentrations lie within a range of about 0.5-3 mg/l across more than two orders of magnitude changes in discharge. In contrast, the warm season data show a high degree of variation based on discharge. At the highest discharges they appear to be very similar to cold season values. At an intermediate range (100-500 m^3/s) they tend to be just slightly lower than cold season values. However, at low discharge values (15-100 m^3/s) the concentrations of NO_3 in the warm season are much lower than the concentrations for these Q values in the cold season, and much more variable (in log units) than they are in either season for any Q values.

The seasonal difference seen in Figure 4 is common among many NO_3 datasets (including many not included in this study). Research on the hydrologic fate and transport of nitrogen suggests a number of reasons why, even in rivers that are often highly enriched in nitrogen, at the lowest flows during the warmer times of the year the NO_3 concentrations can be quite low. When precipitation is low and plants are consuming much of the water in the soil and shallowest parts of the groundwater system, the source of most of the water in the river may be from the deeper parts of the regional groundwater system where the water is very low in NO_3. It may be low in NO_3 either because the water was recharged before the human inputs of nitrogen became large in the last few decades or because the water has had a longer time to denitrify on its longer and slower passage through the groundwater system. In contrast, the water in the river during higher base-flow conditions during the cold season when plants are using much less water could be more associated with the shallow groundwater that has been more influenced by modern inputs of nitrogen. In addition, denitrification taking place in the soil, in groundwater, or at the streambed is much more effective at higher temperatures. Low discharge, which tends to occur mostly in the warmer months, will have a much higher ratio of bed surface area to streamflow volume, and because denitrification is a process that largely occurs on solid surfaces, it will be much more effective at these low flows under warm conditions than under cold conditions. Another important mechanism that leads to this seasonal difference is that the terrestrial and aquatic biological processes that result in uptake of nitrogen from water are much more active in the warm season, so that when streamflow is relatively low, these processes will bring about a much greater removal of nitrate than they would in cold conditions. This general pattern of NO_3 concentrations in relation to streamflow and temperature has been documented in a number of studies such as Fenelon and Moore (1998), Böhlke *et al.* (2007), Alexander *et al.* (2009), and Böhlke *et al.* (2009). These studies document how important streambed denitrification can be under low base-flow conditions particularly at higher temperatures, and that under high base-flow conditions, particularly at lower temperatures, denitrification has a much smaller influence and thus the inputs of high-nitrate shallow groundwater can be a very important determinant of stream NO_3 values.

The way that these seasonal differences are modeled by the L5 or L7 model *vs.* the WRTDS model can have a large impact on flux estimation. This is illustrated in Figure 5. With any of the three models, one can describe the ln(c) *vs.* ln(Q) on any given day of the record as a single curve, simply by substituting the decimal year value for that day into the fitted model. The left panel shows the estimated NO_3 concentrations as a function of discharge when the WRTDS model is used. It presents it for two dates: 2007-02-01 (which is the middle of the "cold season" in an arbitrarily selected year of the dataset) and 2007-08-01 (which is the middle of the "warm season" in the same year). These two curves closely approximate what is evident in the scatter plot shown in Figure 4. The cold season curve is virtually a straight horizontal line (at about 1.8 mg/l). In contrast, the warm season curve has a great deal of curvature in addition to the fact that it is lower than the cold season curve at all discharges. For example, at 45 m^3/s (the 10th percentile on the flow duration curve) the WRTDS August 1 estimate is 0.6 mg/l but at 650 m^3/s (the 90th percentile flow) the WRTDS estimate is 1.3 mg/l. The right panel of Figure 5 shows the estimated NO_3 concentrations on these same two dates

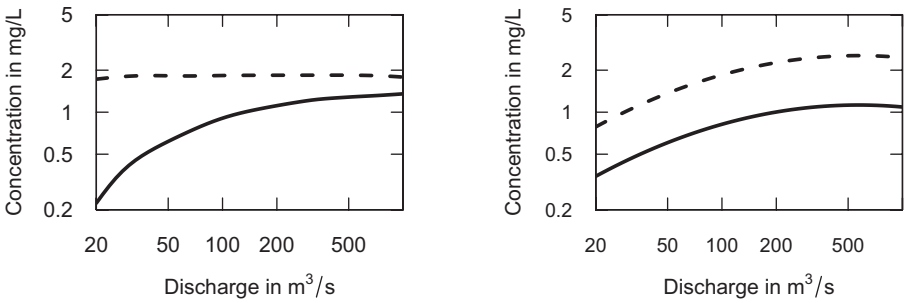

FIGURE 5. Muskingum River at McConnelsville, Ohio, NO_3 Data. Left panel is the WRTDS model estimates for 2007-02-01 (dashed line) and 2007-08-01 (solid line). Right panel is the L7 model estimates for 2007-02-01 (dashed line) and 2007-08-01 (solid line).

as a function of discharge when the L7 model is used. Note that the two curves for the L7 model are exactly parallel to each other, as required by the L7 functional form. For the warm season at 45 m^3/s the estimated concentration is 0.6 mg/l (essentially equal to the WRTDS estimate) but at 650 m^3/s the estimated concentration is 1.1 mg/l (about 15% lower than the WRTDS estimate). But, for the cold season, the estimate at 45 m^3/s is 1.3 mg/l (28% lower than the WRTDS estimate) but for 650 m^3/s the L7 estimate is 2.5 mg/l (38% higher than the WRTDS estimate).

What is the practical significance of these differences when it comes to making long-term average flux estimates? What matters most is how the two methods estimate concentrations at the higher discharges. The much higher concentration estimates for cold season high discharge conditions are what make the greatest difference in the accuracy of the average flux estimate. The requirement that L7 or L5 impose, forcing these curves to be parallel causes the estimates to be too high at the highest discharges in the cold season. It is impossible to isolate the impact of this particular fitting issue from the other two issues. However, it is a problem that is common and has a substantial impact on the bias of flux estimates.

These issues are not limited to NO_3. Richards (2004) noted similar problems with suspended solids in the Maumee River and Dolan and Richards (2008) have noted similar problems for the analysis of TP in five tributaries of Lake Erie and showed that failure to consider these seasonal differences results in significant underestimation of annual fluxes. Of the six cases examined in this study, only two do not appear to have a substantial problem of different shaped $\ln(c)$ *vs.* $\ln(Q)$ relationships in different seasons, those are Maumee TP and Cuya NO_3. In two cases, the WRTDS fitted curves for cool and warm seasons actually cross, with the cold season having substantially higher concentrations than the warm season at discharge values around the 10th percentile flow, but at high-flow (90th percentile) warm season concentrations were higher than cold season. This type of behavior (crossing curves) can never be properly represented by the L5 or L7 model because they both require that the curves be parallel. Failure to properly represent this behavior can be an important contributor to flux bias issues.

Heteroscedastic Behavior

All three of the methods considered here share the same basic approach of building a model in which the dependent variable is $\ln(c)$. It has long been recognized (starting with the work of Ferguson, 1986) that simply transforming these estimates back to real space will produce estimates that have a negative bias. All three of the models considered here use a BCF, a multiplier that compensates for the bias induced by the fact that the model is estimated in log space. In all three cases the estimate of concentration $\hat{c}_i$ can be described in this manner:

$$\hat{c}_i = \mathrm{BCF}_i \cdot \exp(\hat{y}_i)$$

where $\hat{y}_i$ is the regression-based estimate of $\ln(c_i)$ as determined from any of the three models (L7, L5, or WRTDS), and BCF_i is the BCF which is specific to the conditions of the particular day being estimated (the specific discharge and date). Although there are differences in terms of the exact mathematical formulation of the BCF, in general the BCF is closely related to the variance of the regression residuals. To a close approximation, the BCF (for large datasets with no censoring) for all three of these models is:

$$\mathrm{BCF}_i = \exp\left(\frac{\mathrm{SE}_i^2}{2}\right)$$

where SE_i is the estimated standard error of the model for the ith observation. The underlying

assumption of the L5 and L7 models is that the errors are homoscedastic, which means that SE_i is virtually constant across all of the days, and as a consequence the BCF values are virtually constant across all of the days. The slight day-to-day variation in the BCF for these models arises from the parameter uncertainty, but for all practical purposes the BCF for all days can be considered to equal

$$\mathrm{BCF} = \exp\left(\frac{\mathrm{SE}^2}{2}\right)$$

where SE is the standard error of the residuals of the fitted model. In contrast, in the WRTDS model the estimate of SE is allowed to freely vary as a function of discharge, year, and season as interpreted through the fitting of the weighted regression model at a large number of points. Thus, for the WRTDS model the BCF varies widely from day to day because the SE of the model varies widely from day to day as a function of the explanatory variables values on that day.

Using the Musk NO_3 dataset as an example, looking at Figures 1E or 4, it is very clear that variability is greater (in terms of ln(c) values) for very low discharges and particularly so for low discharges during the warmer season of the year than it is for moderate to high discharges. Given this heteroscedasticity of the model, the estimates of the BCF across the sample range from 1.022 to 1.860. For comparison, the BCF for the L7 model only ranges from 1.201 to 1.246. Figure 6 shows the BCF values as a function of Q for all observations in the sample dataset for the L7 model and the WRTDS model.

What this figure shows is that for the higher discharge values, above about 300 m^3/s the BCF for WRTDS is quite small, typically about 1.02 but for these same samples the L7 BCF is about 1.24. Even though both models estimate ln(c) values that are rather similar for these higher discharges. In the case of the WRTDS model they are exponentiated and then multiplied by 1.02. In the case of the L7 model they are exponentiated and then multiplied by 1.24. Because a large portion of the estimated mean flux is contributed by these high discharges, this difference in BCF values has a very substantial impact of flux estimates. For discharges between about 50 m^3/s and 200 m^3/s there is a mixture of cases with the WRTDS BCF being either smaller or larger than for L7, and at the lowest discharges the WRTDS BCF is substantially higher than the L7 BCF because the WRTDS model recognizes the very high error variance present at these low discharges.

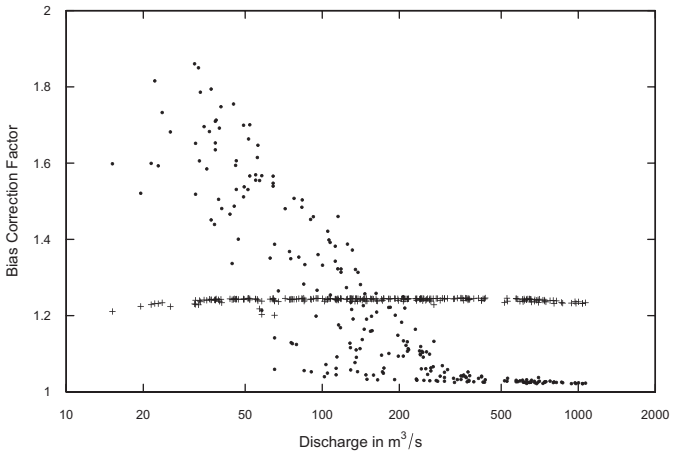

FIGURE 6. Muskingum River at McConnelsville, Ohio, Bias Correction Factor (BCF) *vs.* Discharge, Based on the L7 Model (crosses) and the WRTDS Model (dots).

The differences in the BCF between these models can have a large consequence on annual or long-term flux estimates. In short, the L7 (or L5) model uses a BCF which is much too large for the high Q values and much too small for the low Q values, but because the mean flux estimate is dominated by conditions on the high Q days, the resulting overestimate at high discharges can contribute greatly to the positive bias of long-term mean flux estimates. Of the six cases considered in this study, the heteroscedasticity appears to be problematic in two of them (Musk NO_3 and Rac NO_3) and although there are modest indications of heteroscedasticity in all of the datasets, the variations in variance can be expected to have a small influence on BCF values (e.g., BCF values having differences across models and data points of less than 10%).

Suggested Use of Diagnostic Plots to Identify Potential Bias Problems

Beyond the use of the flux bias statistic B_m, the use of multiple graphical tools can be highly useful to identify and characterize problems with particular model fits for a given dataset. Textbooks on regression methods such as Montgomery *et al.* (2012) strongly encourage the data analyst to use a variety of graphics to identify various problems with the fitted model. What is proposed here is the use of a set of eight graphics that, used together, are likely to elucidate potential bias problems and help the hydrologist diagnose the nature of the problem leading to potential solutions. This set of eight diagnostic graphs is introduced using four specific cases from the resampling experiment. Three of them show serious bias problems (related to one or more of the common causes of bias described above) and one indicates that the model is well suited to the dataset. All of the residuals shown in these graphics, as well as the

estimated values of concentration or flux, use "leave-one-out cross validation" estimates rather than being computed on the full dataset. This approach is used to give a more realistic estimate of prediction errors. This approach to error analysis is described in Appendix S3 in the Supporting Information.

The first example illustrates a very extreme case of bias, which could easily be identified by one or two of these eight plots. Figure 7 shows the application of the L5 to data from the Raccoon River at Des Moines, Iowa. The B_m value is 0.597, in which Figure 3 suggests that it could have a bias greater than 100%. In fact, the true bias E_m for this case is 139%. The discussion will touch on each of the individual panels of the figure.

A. The residuals *vs.* estimated log concentration plot show both a very severe lack of fit as evidenced in the pattern of negative, positive, and then negative residuals moving from low to high estimated log concentrations. Note that these are the residuals in the space in which the model is estimated. That is, they are errors in ln(c). These are the same residuals shown in panels B, C, and D. The estimates are estimates of ln(c). No BCF is involved in the plot shown in panels A, B, C, or D. The residual *vs.* estimated plot for a model that is well suited to the data should show a roughly horizontal cloud of data centered on a residual of zero across the full range of estimated values. In addition to the severe lack of fit, the plot also shows severe heteroscedasticity, with high variance for the low estimated values and lower variance for the high ones. These two features alone should indicate that L5 would be a very poor choice of models. Another type of plot that could be used to identify heteroscedasticity issues is the "Spread-Location plot" which plots the absolute value of the residuals *vs.* the estimated log concentration (see Qian, 2010).
B. The residuals *vs.* log discharge plot looks very much like panel A, and that should be no surprise, given that log discharge is the explanatory variable that explains the greatest portion of the variance in the dataset. It clearly shows lack of fit and heteroscedastic errors. This strong similarity between panels A and B is not always the case.
C. The plot of residuals *vs.* time does not show a lack of fit with respect to time, although it suggests a high degree of serial correlation in the residuals. This serial correlation is not a particularly important issue with respect to bias but can be important to the hydrologist in the overall analysis of the dataset. This kind of plot can be particularly useful to identify cases where there may have been an abrupt change in the system during the period of record that might not be captured by the linear time trend term.
D. The boxplots of residuals by month suggests a lack of fit with respect to the seasonal aspect of the model. Of particular note is the pair of months June and July with almost entirely positive residuals followed by the pair August and September with strongly negative residuals. This relates to the inadequacy of the L5 model for representing differences across seasons (discussed above as the second cause of bias). This panel (and the other panels that use boxplots) follows the convention of setting the width of the box proportional to the square root of the sample size. This can be helpful, particularly in panel D to identify cases where there may be very large differences in sample sizes in different months of the year.
E. The three boxplots of concentration are very informative in this case. The first box shows the distribution of concentrations on sampled days and shows that the highest observed value in the dataset was just slightly less than 20 mg/l. The second box shows the L5 estimates for the set of days that are in the sample. The plot shows a maximum estimate of about 70 mg/l and about a dozen that are in excess of 30 mg/l. This is a very strong indicator of a biased model. In fact, an unbiased model would be expected to show a slightly smaller variability than the actual sample because the estimates will tend to regress to the mean. The third box shows the distribution on all days in the period of record. In this case, they run to even slightly higher values than the sample days. In an unbiased model we might expect that the range of estimates over all days might extend slightly farther than the original sample, because the set of all days may include some very extreme conditions that would result in somewhat extreme estimates, but the high estimates in this figure show that the results are truly not credible. This is one more indication that the L5 model should not be used with this dataset because the model assumptions are so strongly violated.
F. The plot of observed concentrations *vs.* estimated concentrations is another way of looking at the same issues that were seen when comparing the first two boxes in panel E. The solid line is a 1:1 line, which would indicate perfect agreement between the observed and estimated concentrations. This figure shows that the very highest estimates on the sampled days (in the range of 60-70 mg/l) take place on days when

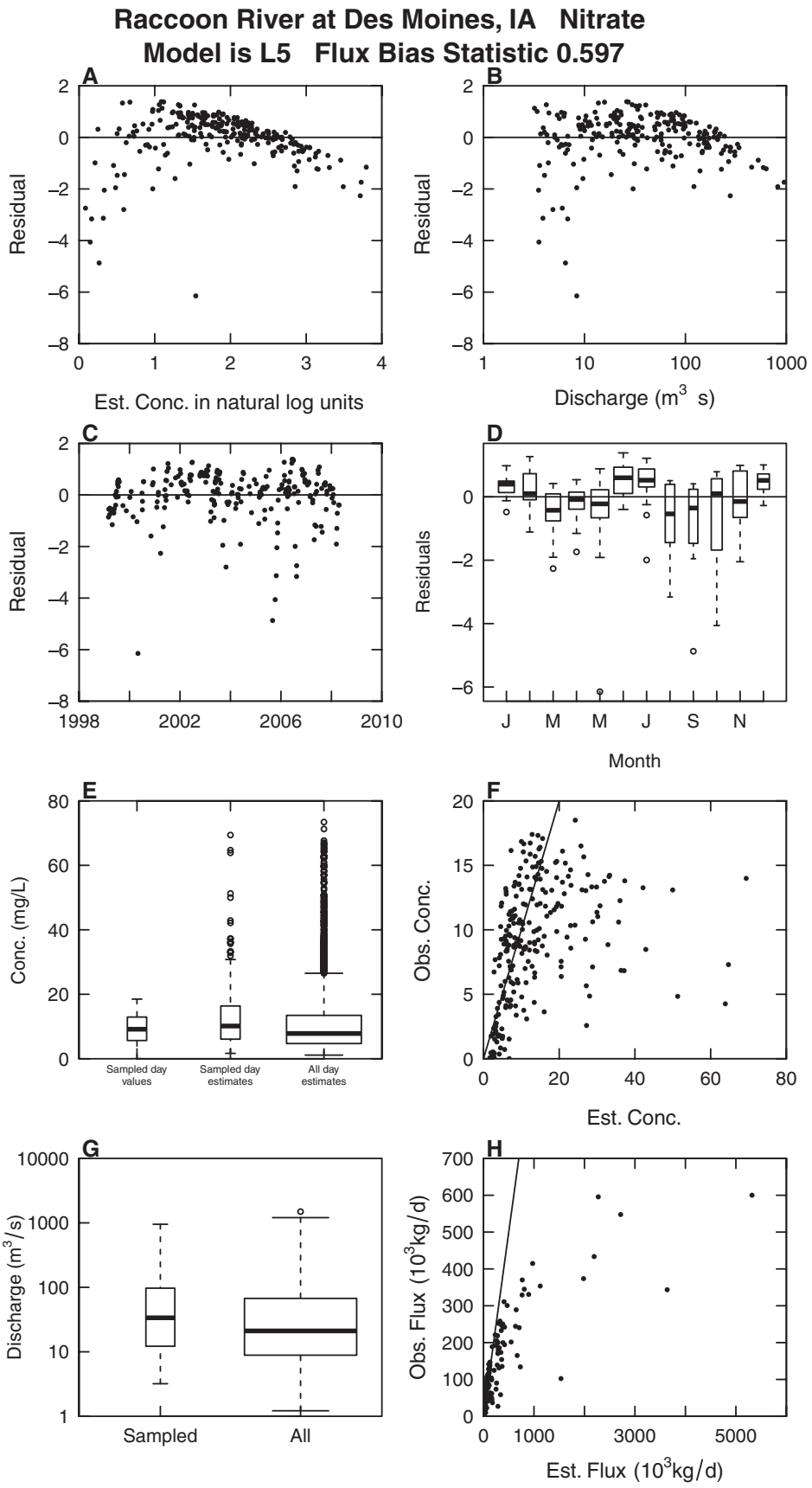

FIGURE 7. Flux Bias Diagnostic Plots for the Raccoon River at Des Moines, Iowa, NO_3, Using the L5 Model.

the true values cover a range from about 4 to 14 mg/l. It is interesting that for some of the observed concentrations the estimates are not far from correct, as seen by their proximity (in the vertical direction) to the 1:1 line. But, the overall severe lack of symmetry in terms of points lying above and below the 1:1 line is another strong indicator of bias.

G. The pair of boxplots showing discharge (on a log scale) for the sample dataset and discharge for the full set of days in the period of record is not intended to be an indicator of a fitting problem. Rather, this plot is included to alert the hydrologist to differences between the distribution of daily mean discharges on sampled days, and the daily mean discharge on all days in the period of record. Ideally, one would like to see the sampled days distribution to be shifted slightly upwards compared to the distribution of all days (as is the case here). If one were to observe the opposite, where the distribution of sampled days is displaced somewhat lower than all days, this would be a cause for concern. If the highest discharges (those well above the upper quartile) were rarely sampled then it would be nearly impossible to determine if any flux bias existed. In this case the set of samples is quite adequate.

H. Finally, the plot of observed flux *vs.* estimated flux (shown here with a 1:1 line) is the most simple and direct representation of the flux bias issue. By making the plot on an arithmetic scale, many observations cluster together near the origin and thus are rather obscure. The emphasis of the figure is on the accuracy of estimates on days of high observed and/or estimated flux. What is important in this example is the great lack of symmetry around the 1:1 line. The figure shows estimated flux values in the range of 2,000,000 kg/day to nearly 6,000,000 kg/day when the actual fluxes on those days are generally in the range of 400,000-600,000 kg/day. This plot provides very decisive evidence of flux bias.

Figure 8 shows same set of diagnostic graphics based on the L7 fit to these data. The B_m statistic for this model is 0.319, a good deal smaller than with the L5 model, but still indicative of a serious positive bias. The actual bias in this case was 54%. Panel A shows the same basic problem of lack of fit seen in Figure 7. Panel B shows that the quadratic term does help to resolve problems due to the curvature of the ln(c) *vs.* ln(Q) relationship. Panel D shows the poor fit of the seasonal pattern. Panels A and B also show heteroscedasticity similar to the L5 model (seen in Figure 7). Panels E, F, and H all show the extreme overestimation of the upper range of estimated concentrations. The overestimation at higher discharge values is less pronounced than in the L5 case but it is still very extreme.

Figure 9 shows the diagnostic graphics for the WRTDS fit for this same dataset. In this case, the B_m statistic is virtually zero (−0.00237) and the actual bias is −4%. Panels A and B both show that the errors are heteroscedastic, but there is no particular lack of fit overall or with respect to discharge, time, or season. The pattern of estimated concentrations or fluxes is highly consistent with the observed patterns. The heteroscedasticity shown here is fundamental to the data and does not present a problem for the model, because no assumption of homoscedasticity is made in the WRTDS model. Panel E shows that the sample data are just slightly more variable than the estimates (either for the sampled days or all days). This is exactly the type of behavior one should expect, with estimates regressing somewhat toward the mean. Finally, panels F and H both show a roughly symmetrical pattern of error surrounding the 1:1 line. Indicating that for concentration and for flux the results have little bias.

One final example of the flux bias diagnostic plots is shown in Figure 10. This is the Honey TP dataset using the L7 model. In this case the B_m statistic is 0.33, indicative of a rather strong positive bias. The actual bias in this case is 51%. Panel A of Figure 10 shows the severe lack of fit, with all 10 of the highest predicted values having negative residuals, and more moderate estimates having strongly positive residuals. Panel E shows estimates over all days reaching values as high as 3.4 mg/l even though observed concentrations never exceeded 1.5 mg/l. Panel H shows the four highest estimated flux values on sampled days were between about 8,000 and 10,000 kg/day. These days actually had flux values in the 2,000-4,500 kg/day range. In short, the lack of fit (which is the primary issue in this case) can be readily seen in panels A, B, E, F, and H.

Overall, the use of diagnostic graphics in evaluating flux estimation methods is highly beneficial. Arguably they may be viewed as "overkill" but they can be helpful by providing multiple perspectives on possible problems with the use of a given model for the dataset. The plots can be an automatic output from statistical software used for flux estimation and can quickly aid the hydrologist in identifying the potential for severe bias problems. (They are already included in the EGRET software package and can be used to examine estimates made by any model.) They may also be very helpful for identifying coding errors in a dataset or for identifying patterns that can be very informative about sources of variability that might show up as anomalies in the plot of residuals

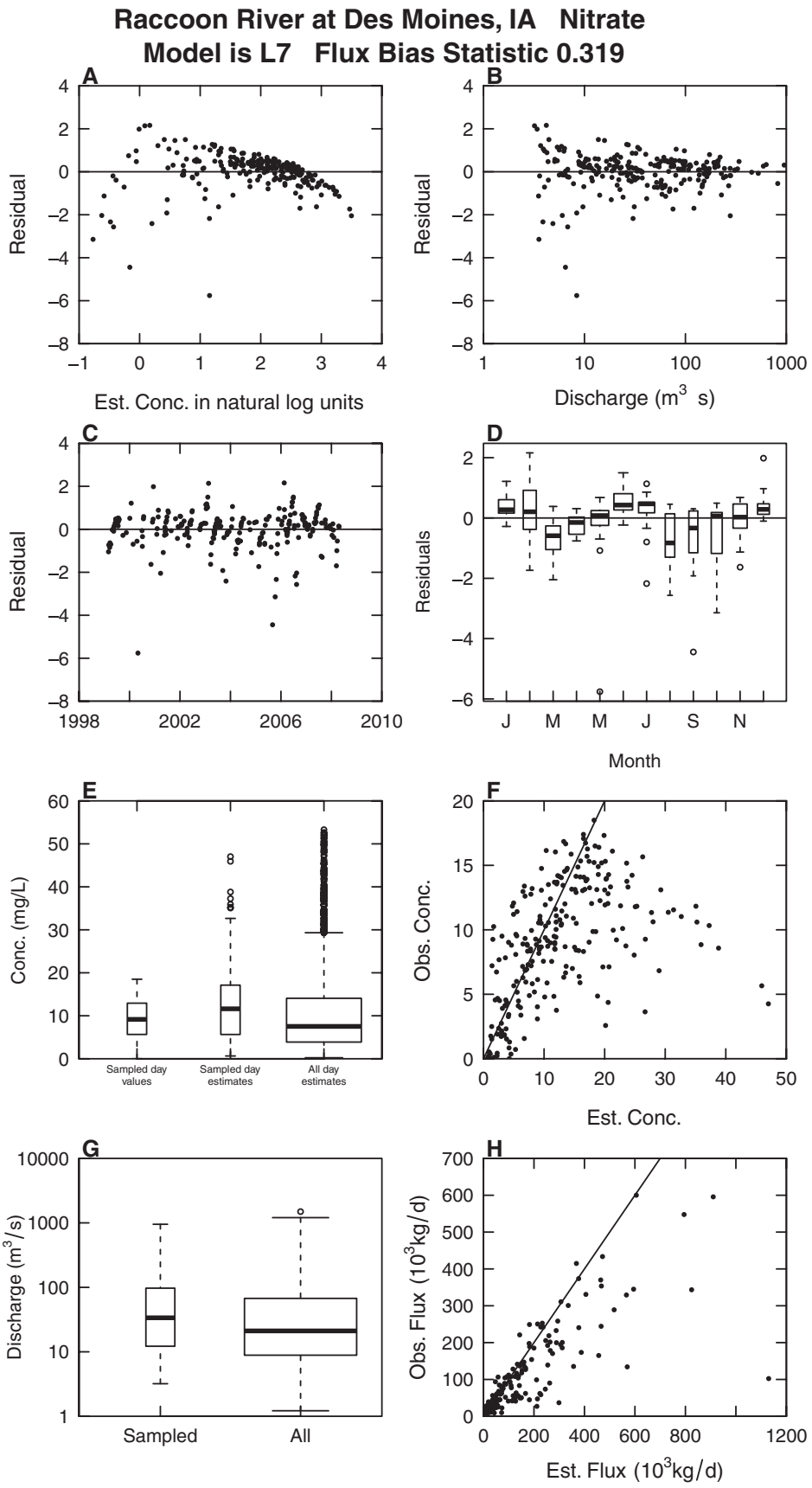

FIGURE 8. Flux Bias Diagnostic Plots for the Raccoon River at Des Moines, Iowa, NO_3, Using the L7 Model.

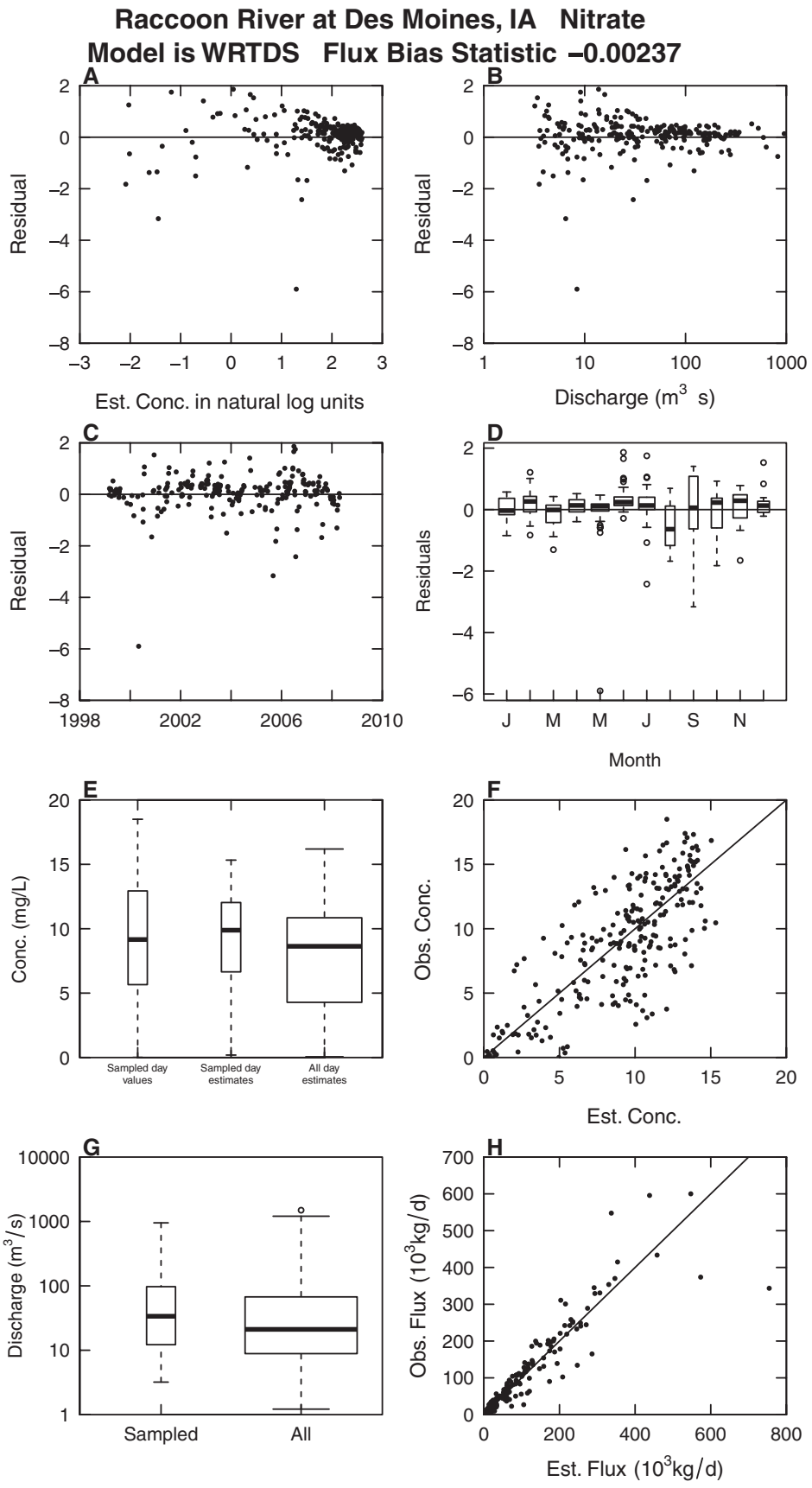

FIGURE 9. Flux Bias Diagnostic Plots for the Raccoon River at Des Moines, Iowa, NO_3, Using the WRTDS Model.

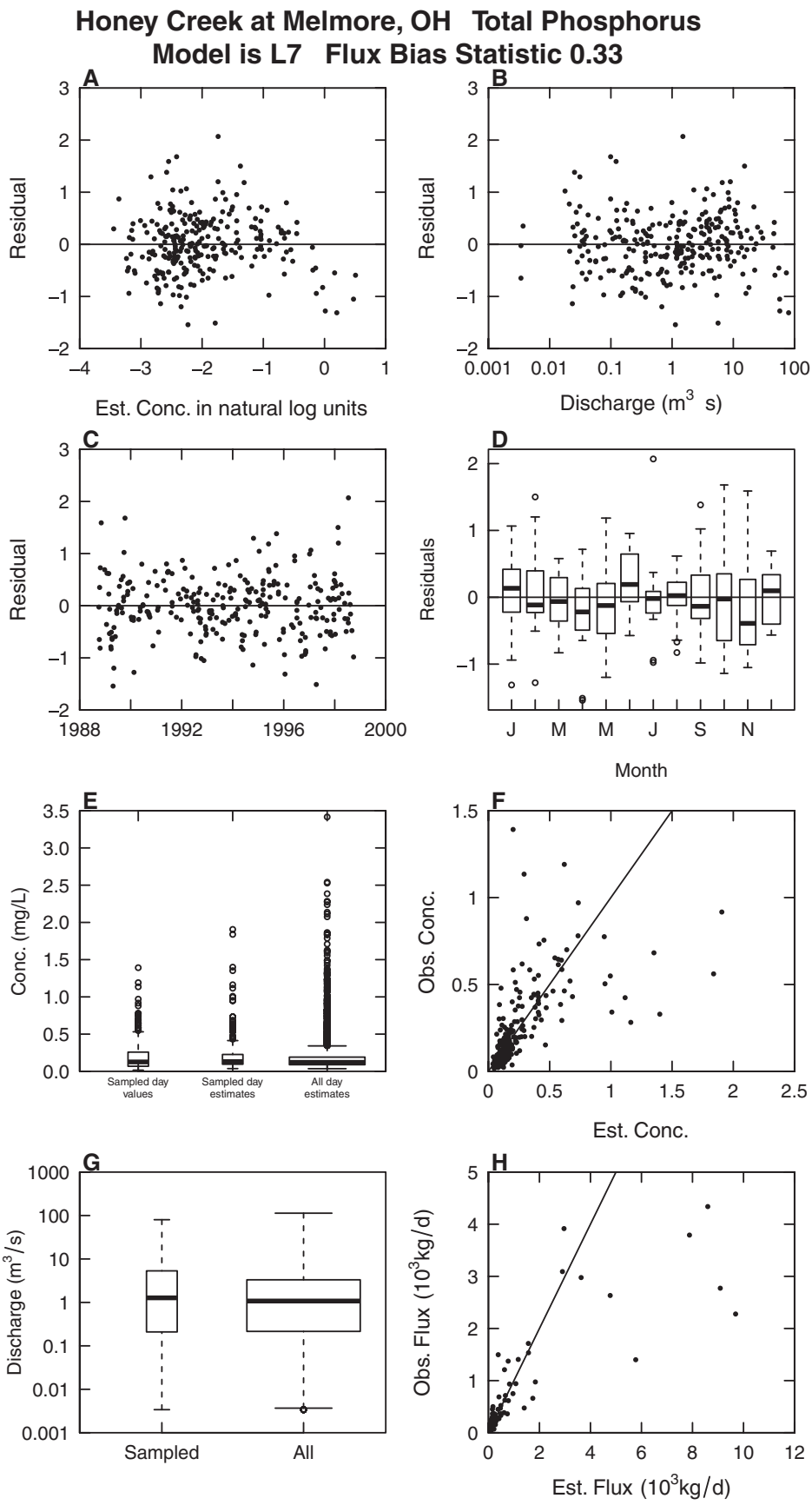

FIGURE 10. Flux Bias Diagnostic Plots for Honey Creek at Melmore, Ohio, Using the L7 Model.

as a time series or the boxplots of residuals by month. The collection of plots in conjunction with the flux bias statistic does not provide a "yes or no" switch about using a particular model for a given dataset. Rather, they serve as an aid to the hydrologist in identifying issues that need to be considered and can lead directly to the identification of some of the more common problems.

DEALING WITH RESULTS WITH A KNOWN FLUX BIAS PROBLEM

One common approach to resolving a problem of flux bias is to include additional terms in the regression model that may help to improve the overall quality of the model and reduce the bias. If another variable is able to substantially improve the overall model fit, particularly for the high-flow cases that contribute so much to the bias problem, then modifying the regression model to include such variables can be very helpful. The LOADEST model as implemented by Runkel *et al.* (2004) is designed to accept such additional explanatory variables. One example of this approach is described in Garrett (2012), which examines fluxes of a number of constituents in 10 major rivers within the state of Iowa for the period 2004-2008. The approach used there (described on pages 7-8 of Garrett (2012)) was to consider a statistic (functionally equivalent to B_m) and to reject those models for which B_m was deemed to be very different from the ideal value of zero. Model selection used a combination of residuals plots (such as those used here) and the Akaike Information Criteria. Other variables that were used to improve the fit of these models included the following: a term that considers hysteresis (Wang and Linker, 2008) based on the change over the past day or the past 30 days, variables that allow the $\ln(c)$ *vs.* $\ln(Q)$ relationship to be piecewise linear with a total of two pieces, and the use of discharge anomaly terms that characterize antecedent conditions at the time of sampling (Vecchia *et al.*, 2009). These anomaly terms are designed to integrate flow conditions for periods as short as a few days to as much as a year or longer. The idea is that concentration may be strongly related both to current discharge (on the sampling day) and discharge over some antecedent period. This approach has been used in conjunction with models such as L5 and L7. It has not been implemented in conjunction with a WRTDS approach, but in data-rich settings this may have a good deal of potential for improving flux estimates. In a previous study, an additional variable based on the time elapsed since the most recent hydrograph rise was shown to improve on the estimates from simpler regression models (Hirsch, 1988).

In summary, a wide range of experiences has shown that simply using a single functional form, such as L5 or L7, across a wide range of cases can result in a mixture of unbiased and severely biased results. Adding additional variables where needed, such as suggested in these studies, or adding flexibility to the functional form of the model, such as WRTDS, can often provide relatively unbiased estimates of long-term flux in a large number of cases.

TOPICS IN NEED OF FURTHER EXPLORATION

There are a number of additional questions that should be examined in further research on the flux bias issue.

Flux estimation with much smaller datasets presents significant challenges. This analysis only considers datasets of at least 120 observations. Diagnosis of the types of problems considered here becomes very difficult with small datasets, and addition of more explanatory variables is problematic for small datasets.

The resampling experiment results shown in this study were all designed to provide a set of water quality samples collected at a set of discharges that follow a probability distribution that is approximately the same as the distribution of the full set of daily discharge data, or in the stratified random cases, intentionally oversamples the higher end of the discharge distribution. Some sampling strategies are known to substantially undersample the high discharges, or not sample them at all. These cases would undoubtedly be more prone to large errors and large biases than those presented in this study. Even more caution is needed in these cases than is the case with the random or stratified random sampling cases considered here. Finding diagnostics for such cases remains a challenge.

One strategy for dealing with small sample sizes and/or poor representation of high-flow samples may be the use of Bayesian approaches to estimation. Such an approach would use experience from richer records at sites that are likely to be relevant to the site at hand. It should be possible to build a prior probability model for the first and second derivatives of the $\ln(c)$ *vs.* $\ln(Q)$ relationship from a collection of relatively rich datasets. Using this prior with the information from a specific smaller dataset can help to avoid unreasonable functional forms that may

result from classical regression analysis. Bayesian methods would assure that when at-site information is relatively rich, it provides most of the information that shapes the model, but when the at-site information is relatively poor, the regionally based prior model exerts the dominant influence. It may also be possible to use short periods of highly intense sampling information (from *in situ* sensors) for a short period of time to help form a prior regarding the shape of the function, even if the data cover only a small portion of the total period of record that is of interest.

The issue of censored data is not considered in this study. Limited experimentation indicates that a moderate amount of censoring does not change the overall bias problem. The issue of how censoring affects bias needs to be explored. Doing that exploration is complex and probably needs to be considered through experiments that use artificial censoring. Consideration of various degrees of censoring adds an additional dimension of complexity to the experimental design, but the experiments are certainly worth undertaking. The presence of censoring also creates challenges for producing useful versions of diagnostic plots such as those shown in Figures 7-10.

Finally, there has been no comparable evaluation of the bias associated with nonregression-based estimators. These include estimates based on interpolation of concentrations in the time domain and also a variety of ratio estimator methods. In cases where relationships between discharge and concentration are weak or highly complex these methods may be very worthwhile alternatives to be explored, but little is known about their biases.

CONCLUSIONS

Estimation of long-term fluxes for a wide range of substances is crucial to the understanding of water quality and underpins the management of many water quality issues. Because of the difficulty and expense of data collection, most sites of interest have only a relatively sparse set of samples (say a dozen to a few dozen per year) of the substance of interest and these are used along with long-term daily discharge data to estimate annual or long-term mean fluxes. It has been widely recognized that some of the standard models of doing such calculations are vulnerable to large systematic biases. This study focuses on understanding the origins and potential magnitudes of flux-bias problems associated with three common regression-based models of estimation. Experience and a wide range of unpublished experimentation has shown that most or all of these models often produce very nearly unbiased estimates of flux, but in some specific types of situations the estimates that one or more of them produce can be severely biased.

These large biases arise for one or more of the following reasons: (1) there is a severe lack of fit in terms of the $\ln(c)$ *vs.* $\ln(Q)$ relationship, (2) this relationship has a substantially different shape in different seasons of the years, and (3) the errors from the fitted model are severely heteroscedastic. The results of this study suggest that of the three models considered, WRTDS represents a more robust approach to flux estimation, generally exhibiting biases that are no worse than the L5 or L7 models, and in a number of cases exhibiting biases that are much smaller than those from L5 or L7.

These results suggest that a robust approach would be to use WRTDS in cases with sample sizes of 120 observations or more. However, this does not mean that selecting WRTDS assures a nearly unbiased result, but they do suggest that there is a smaller chance of severe bias. The use of diagnostic plots and the flux bias statistic, B_m, can be very useful tools for indicating the possibility of severe biases and can help guide the hydrologist to build and evaluate alternative models (including ones that are more complex than those considered here). There is no universal "right way" to compute unbiased estimates of mean flux, but the use of a more robust approach and a set of diagnostic tools can be very useful in limiting these errors.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Data preparation and determination of "true flux."

Appendix S2. Stratified random sampling experiment.

Appendix S3. Use of leave-one-out cross validation residuals.

LITERATURE CITED

Alameddine, I., S.S. Qian, and K.H. Reckhow, 2011. A Bayesian Changepoint-Threshold Model to Examine the Effect of TMDL Implementation on the Flow-Nitrogen Concentration Relationship in the Neuse River Basin. Water Research 45:51-62, doi: 10.1016/j.watres.2010.08.003.

Alexander, R.B., J.K. Böhlke, E.W. Boyer, M. David, J.W. Harvey, P.J. Mulholland, S.P. Seitzinger, C.R. Tobias, C. Tonitto, and W.M. Wollheim, 2009. Simulating Temporal Variations of Nitro-

gen Losses in River Networks with a Dynamic Transport Model Unravels the Coupled Effects of Hydrological and Biogeochemical Processes. Biogeochemistry 93:91-116. http://www.springerlink.com/content/441650023u7k58u9/, *accessed* August 2013.

Böhlke, J.K., R.C. Antweiler, J.W. Harvey, A.E. Laursen, L.K. Smith, R.L. Smith, and M.A. Voytek, 2009. Multi-Scale Measurements and Modeling of Denitrification in Streams with Varying Flow and Nitrate Concentration in the Upper Mississippi River Basin, USA. Biogeochemistry 93:117-141. http://www.springerlink.com/content/f00r657809770131/, *accessed* August 2013.

Böhlke, J.K., M.E. O'Connell, and K.L. Prestegaard, 2007. Ground-Water Stratification and Delivery of Nitrate to an Incised Stream in Varying Flow Conditions. Journal of Environmental Quality 36:664-680. https://dl.sciencesocieties.org/publications/jeq/abstracts/36/3/664, *accessed* August 2013.

Cohn, T.A., 2005. Estimating Contaminant Loads in Rivers: An Application of Adjusted Maximum Likelihood to Type 1 Censored Data. Water Resources Research 41:W07003, doi: 10.1029/2004WR003833, 13 pp.

Cohn, T.A., D.L. Caulder, E.J. Gilroy, L.D. Zynjuk, and R.M. Summers, 1992. The Validity of a Simple Statistical Model for Estimating Fluvial Constituent Loads: An Empirical Study Involving Nutrient Loads Entering Chesapeake Bay. Water Resources Research 28(9):2353-2363.

Cohn, T.A., L.L. DeLong, E.J. Gilroy, R.M. Hirsch, and D.K. Wells, 1989. Estimating Constituent Loads. Water Resources Research 25(5):937-942.

Crawford, C.G., 1991. Estimation of Suspended-Sediment Rating Curves and Mean Suspended-Sediment Loads. Journal of Hydrology 129:331-348.

Crowder, D.W., M. Demissie, and M. Markus, 2007. The Accuracy of Sediment Loads When Log-Transformation Produces Nonlinear Sediment Load-Discharge Relationships. Journal of Hydrology 336:250-268, doi: 10.1016/j.jhydrol.2006.12.024.

Dolan, D.M. and R.P. Richards, 2008. Analysis of Late 90s Phosphorus Loading Pulse to Lake Erie. *In*: Checking the Pulse of Lake Erie, M. Munawar and R. Heath (Editors). Ecovision World Monograph Series, Schweizerbart Science Publishers, Stuttgart, Germany, pp. 79-96.

Dolan, D.M., A.K. Yui, and R.D. Geist, 1981. Evaluation of River Load Estimation for Total Phosphorus. Journal of Great Lakes Research 7(3):207-214.

Fenelon, J.M. and R.C. Moore, 1998. Transport of Agrichemicals to Ground and Surface Water in a Small Central Indiana Watershed. Journal of Environmental Quality 27:884-894.

Ferguson, R.I., 1986. River Loads Underestimated by Rating Curves. Water Resources Research 22(1):74-76.

Ferguson, R.I., 1987. Accuracy and Precision of Methods for Estimating River Loads. Earth Surface Processes and Landforms 12(1):95-104.

Garrett, J.D., 2012. Concentrations, Loads, and Yields of Selected Constituents from Major Tributaries of the Mississippi and Missouri Rivers in Iowa, Water Years 2004-2008. U.S. Geological Survey Scientific Investigations Report 2012-5240, 61 pp. http://pubs.usgs.gov/sir/2012/5240/.

Guo, Y., M. Markus, and M. Demissie, 2002. Uncertainty of Nitrate-N Load Computations for Agricultural Watersheds. Water Resources Research 38(10):1185, doi: 10.1029/2001WR001149.

Hirsch, R.M., 1988. Statistical Methods and Sampling Design for Estimating Step Trends in Surface-Water Quality. Water Resources Bulletin 24(3):493-503.

Hirsch, R.M., 2012. Flux of Nitrogen, Phosphorus, and Suspended Sediment from the Susquehanna River Basin to the Chesapeake Bay during Tropical Storm Lee, September 2011, as an Indicator of the Effects of Reservoir Sedimentation on Water Quality. U.S. Geological Survey Scientific Investigations Report 2012-5185, 17 pp. http://pubs.usgs.gov/sir/2012/5185/.

Hirsch, R.M., D.L. Moyer, and S.A. Archfield, 2010. Weighted Regressions on Time, Discharge, and Season (WRTDS), with an Application to Chesapeake Bay River Inputs. Journal of the American Water Resources Association 46(5):857-880, doi: 10.1111/j.1752-1688.2010.00482.x.

Medalie, L., R.M. Hirsch, and S.A. Archfield, 2012. Use of Flow-Normalization to Evaluate Nutrient Concentration and Flux Changes in Lake Champlain Tributaries, 1990-2009. Journal of Great Lakes Research 38(Suppl. 1):58-67. http://dx.doi.org/10.1016/j.jglr.2011.10.002, *accessed* August 2013.

Montgomery, D.C., E.A. Peck, and G.G. Vining, 2012. Introduction to Linear Regression Analysis (Fifth Edition). John Wiley and Sons, Hoboken, New Jersey, 672 pp.

Moyer, D.L., R.M. Hirsch, and K.E. Hyer, 2012. Comparison of Two Regression-Based Approaches for Determining Nutrient and Sediment Fluxes and Trends in the Chesapeake Bay Watershed. U.S. Geological Survey Scientific Investigations Report 2012-5244, 118 pp. http://pubs.usgs.gov/sir/2012/5244.

Preston, S.D., R.B. Alexander, G.E. Schwarz, and C.G. Crawford, 2011. Factors Affecting Stream Nutrient Loads: A Synthesis of Regional SPARROW Model Results for the Continental United States. Journal of the American Water Resources Association 47(5):891-915, doi: 10.1111/j.1752-1688.2011.00577.x/full.

Preston, S.D., V.J. Bierman, Jr., and S.E. Sillman, 1989. An Evaluation of Methods for the Estimation of Tributary Mass Loads. Water Resources Research 25:1379-1389.

Qian, S.S., 2010. Environmental and Ecological Statistics with R. Chapman and Hall/CRC Press, Boca Raton, Florida, 421 pp.

Richards, R.P., 2004. Improving TMDLs with Lessons Learned from Long-Term Detailed Monitoring. Journal of Environmental Engineering 130:657-663.

Richards, R.P., I. Alameddine, J.D. Allan, D.B. Baker, N.S. Bosch, R. Confesor, J.V. DePinto, D.M. Dolan, J.M. Reutter, and D. Scavia, 2012. Discussion of "Nutrient Inputs to the Laurential Great Lakes by Source and Watershed Estimated Using SPARROW Watershed Models" by Dale M. Robertson, and David Saad. Journal of the American Water Resources Association 49(3):715-724, doi: 10.1111/jawr.12006/full, downloaded March 4, 2013.

Richards, R.P. and J. Holloway, 1987. Monte Carlo Studies of Sampling Strategies for Estimating Tributary Loads. Water Resources Research 23:1939-1948.

Robertson, D.M. and E.D. Roerish, 1999. Influences of Various Water Quality Sampling Strategies on Load Estimates for Small Streams. Water Resources Research 35(12):3747-3759.

Runkel, R.L., C.G. Crawford, and T.A. Cohn, 2004. Load Estimator (LOADEST) – A FORTRAN Program for Estimating Constituent Loads in Streams and Rivers. U.S. Geological Survey Techniques and Methods, Book 4, Chap. A5, 75 pp. http://pubs.er.usgs.gov/publication/tm4A5, *accessed* August 2013.

Smith, R.A., G.E. Schwarz, and R.B. Alexander, 1997. Regional Interpretation of Water-Quality Monitoring Data. Water Resources Research 33:2781-2798.

Sprague, L.A., R.M. Hirsch, and B.T. Aulenbach, 2011. Nitrate in the Mississippi River and Its Tributaries, 1980-2008: Are We Making Progress? Environmental Science and Technology 45(17):7209-7216, doi: 10.1021/es201221s/, *accessed* August 2013.

Stenback, G.A., W.G. Crumpton, K.E. Schilling, and M.J. Helmers, 2011. Rating Curve Estimation of Nutrient Loads in Iowa Rivers. Journal of Hydrology 396:158-169, doi: 10.1016/j.jhydrol.2010.11.006.

Tobin, J., 1958. Estimation of Relationships for Limited Dependent Variables. Econometrica 26:24-36.

Vecchia, A.V., R.J. Gilliom, D.J. Sullivan, D.L. Lorenz, and J.D. Martin, 2009. Trends in Concentrations and Use of Agricultural

Herbicides for Corn Belt Rivers, 1996-2006. Environmental Science and Technology 43(24):9096-9102.

Verma, S., M. Markus, and R.A. Cooke, 2012. Development of Error Correction Techniques for Nitrate-N Load Estimation Methods. Journal of Hydrology 12–25:432-433, doi: 10.1016/j.jhydrol.2012.02.011.

Wang, P. and L.C. Linker, 2008. Improvement of Regression Simulation in Fluvial Sediment Loads. Journal of Hydraulic Engineering 134(10):1527-1531.

Zhang, Q., D.C. Brady, and W.P. Ball, 2013. Long-Term Seasonal Trends of Nitrogen, Phosphorus, and Suspended Sediment Load from the Non-Tidal Susquehanna River Basin to Chesapeake Bay. Science of the Total Environment 452(3):208-221, doi: 10.1016/j.scitotenv.2013.02.012.

Large biases in regression-based constituent flux estimates: causes and diagnostic tools

Robert M. Hirsch[1]

Appendix S1

Data preparation and determination of "true flux"

As described above the first step is to determine the "true flux" for as many days as possible for the available data set. Define $F_{i,j}$ to be the "true flux" for day i, of year j. Of course "true flux" is never known perfectly, but the cases presented here are ones in which the sampling is so intensive that errors are expected to be rather small. The method of determining $F_{i,j}$ can be summarized as follows.

1. If there are no samples on a given day, then $F_{i,j}$ is considered to be a missing value.
2. If there is one or more sample on day i of year j, then $F_{i,j}$ is computed as the estimated average flux over that day. In computing this average the assumption is made that concentration changes as a step function, with the steps located at the mid-way point between sampling times. The samples considered for any given day include the last sample taken prior to that day and the first sample taken after the day. Concentration for each segment of the day is assumed to be a constant, which is equal to the concentration from the sample that was collected during this segment of the day. Thus, no "rating curve" relationship is used to estimate concentrations between samples.
3. Estimated discharge is determined for each of the time segments for which there is a concentration estimate. If there is no recorded estimate of discharge at the time of sampling, then the discharge for all segments of the day is equal to the published daily mean discharge for the USGS streamgage. If there is a recorded estimate of discharge for each time of sampling (which is the case for the Heidelberg University data) then the discharge for each time segment of the day is based on the discharge associated with the sample taken during that segment. However, one final adjustment is made. The discharges for each segment of the day are multiplied by an adjustment factor to assure that the average discharge over all segments of the day equal the published daily mean discharge from the USGS streamgage. This final step is necessary because the discharge estimates in the Heidelberg University data sets use discharge values that are derived from the stage value at the time of sampling and a standard (unshifted) rating curve. Thus the recorded discharge values in the Heidelberg records are

[1] Research Hydrologist, U.S. Geological Survey, 432 National Center, USGS, Reston, VA, 20192 (Email: rhirsch@usgs.gov)

likely to be a good indicator of the pattern of discharge change over the course of the day, but not a good basis for determining a true discharge for the day.

For a full description of the data sets collected by the Heidelberg University National Center for Water Quality Research see http://www.heidelberg.edu/academiclife/distinctive/ncwqr. The data sets were downloaded from that site. It is noteworthy that these data sets are often as much as 30 years in duration and may contain more than 18,000 samples. Many dates are sampled multiple times during the course of the day and the time of day and approximate discharges at the sampling time are included in the data set. These sites are all on tributaries to Lake Erie or the Ohio River. The one data set not from the NCWQR is the NO3 record from the Raccoon River in Iowa. It was collected by the Des Moines, Iowa waterworks. These samples are much less dense. 46 percent of the days are sampled and each sampled day has only one sample. For computational purposes the sample is assumed to have been collected at noon each day. The Raccoon River data set was included because it represents an example of a site with very intense agriculture and a pattern of relationships among streamflow, seasons, and concentration that are typical of some of the most agriculturally intensive watersheds in the nation. In all of the cases studied the experiment was applied to a period of about 10 years.

Appendix S2

Stratified random sampling experiment.

In addition to the re-sampling experiment described in the body of the paper, 180 more data sets were selected based on a stratified random sampling strategy. In each of these stratified random sample cases the probability of selecting a particular sample varied with discharge. The full set of samples in the record was divided into three classes: a low-to-moderate flow class (about the 0 to 60th percentile on the flow duration curve), a high flow class (about the 60th to 95th percentile), and a very high flow class (95th percentile or above). The probabilities for selecting a sample from the very high flow class was twice the probability of selecting a sample from the high flow class, and the probability of selecting from this class was twice that for the low-to-moderate flow class. This stratified random scheme is more like the strategies used by the USGS at many long-term monitoring sites, and can be expected to be more accurate and less biased that estimates based on a random sample, because the quality of the fit will be better at the high flows where most of the flux happens. The reason for considering the stratified random sampling case was to determine if its use might substantially mitigate the bias problem.

The results of the stratified random sampling experiment are shown in figure B1. The show that this type of sampling can help to mitigate the bias problem but it does not provide a general cure for the problem.

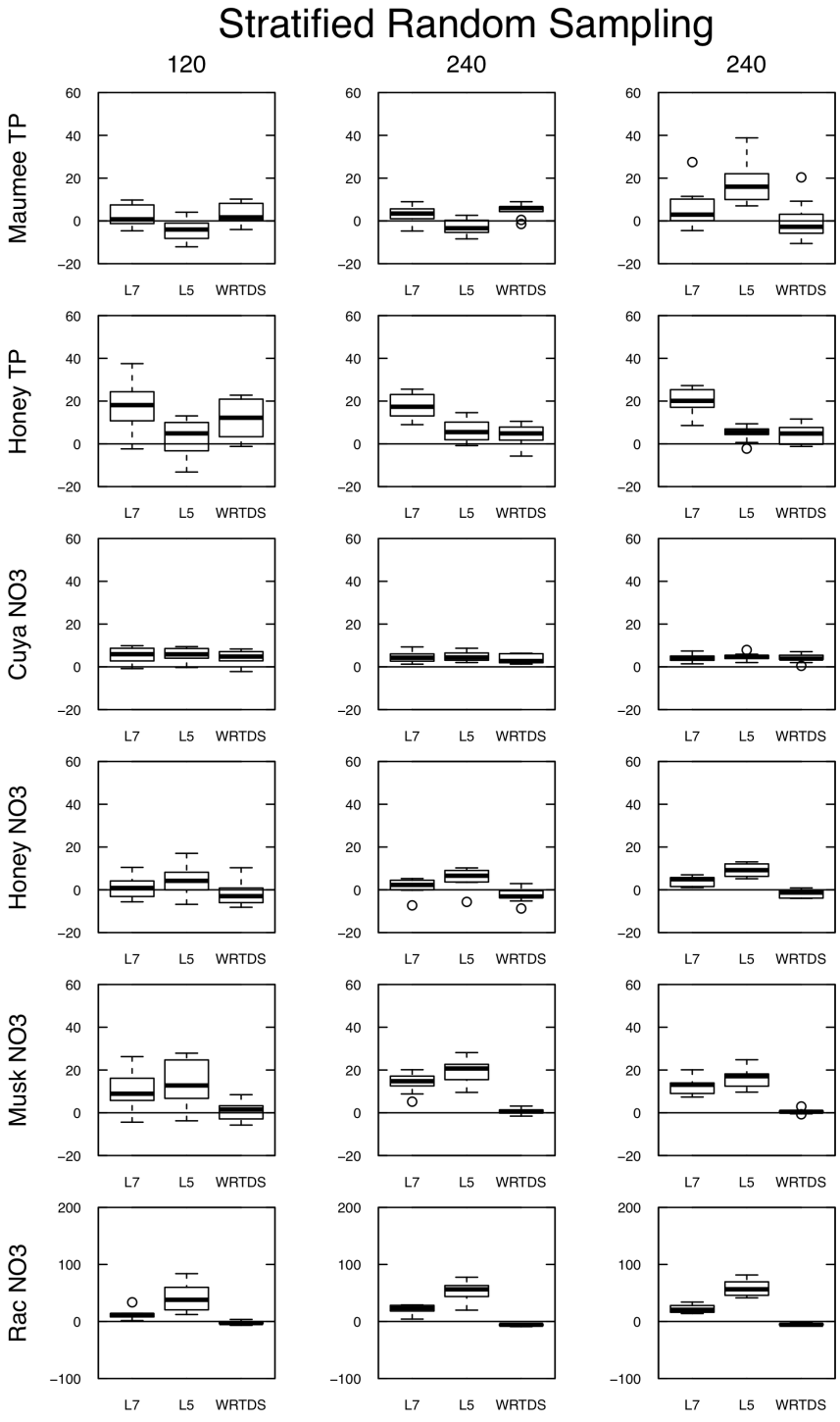

Figure S2-1. Boxplots of the error in percent for estimates of average flux, for six data sets selected using stratified random sampling, at sample sizes of 120, 240 and 480 samples, for three models (L7, L5, and WRTDS). Each box represents the results of 10 subsamples of the full data set. Note the scale differences.

Appendix S3

Use of leave-one-out cross validation residuals.

The three models used in this paper have varying degrees of flexibility. The least flexible is L5, next is L7, and WRTDS is substantially more flexible than either of the others. Given that difference, it is important that the type of residuals analysis being done uses a cross-validation type of approach. In regression analysis, one approach used in model selection is called "leave-one-out cross validation." This involves computing prediction residuals. These prediction residuals (PRs) form the basis for the PRESS statistic (Prediction Error Sums of Squares), a common metric used in selecting among a variety of competing regression models. (Montgomery et al., 2012, p. 134). The prediction residual for any of these models is computed as:

$$e_{(i)} = y_i - \hat{y}_{(i)}$$

$e_{(i)}$ is the ith prediction residual

y_i is the true value of the dependent variable, which in all three models is ln(c_i)

and, $\hat{y}_{(i)}$ is the estimate of the quantity y_i with the ith observation left out of the analysis. The PR should give a more realistic representation of the ability of the model to predict values for days when the true value is not known. Particularly in the case of WRTDS there is a legitimate question as to whether the quality of fit to a specific observation, particular when that observation is extreme in terms of one or more of the explanatory variables, is simply a result of the model "tuning" its coefficients to come close to correctly fitting that one observation. For the L5 and L7 models the prediction residual is easily calculated as

$$e_{(i)} = \frac{e_i}{1 - h_i}$$

where h_i is the leverage of the ith observation. For a definition of leverage, see Montgomery et al. (2012). The leverage statistics on each observation are generally computed in statistical software packages.

In the case of the WRTDS model the prediction residuals must be calculated by re-estimating the model at each of the n points, using the remaining n-1 observations. This calculation is included in the WRTDS implementation contained in the EGRET package in R (https://github.com/USGS-CIDA/WRTDS/wiki). All residuals shown in graphics in this paper are prediction residuals. If sample sizes are large, the difference between the graphics showing the prediction residuals and ordinary residuals will be rather minor, but given the relative ease with which they are computed, and their relevance to actual predictions, they are used in this paper.