Journal of Hydrology 542 (2016) 185–203

Contents lists available at ScienceDirect

Journal of Hydrology

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

Research papers

An evaluation of methods for estimating decadal stream loads

Casey J. Lee[a,*], Robert M. Hirsch[b], Gregory E. Schwarz[b], David J. Holtschlag[c], Stephen D. Preston[d], Charles G. Crawford[e], Aldo V. Vecchia[f]

[a] U.S. Geological Survey, Kansas Water Science Center, 4821 Quail Crest Place, Lawrence, KS 66049, USA
[b] U.S. Geological Survey, 12201 Sunrise Valley Dr., Reston, VA 20192, USA
[c] U.S. Geological Survey, Michigan Water Science Center, 6520 Mercantile Way #5, Lansing, MI 48911, USA
[d] U.S. Geological Survey, 1289 McDaniel Dr., Dover, DE 19901, USA
[e] U.S. Geological Survey, 5957 Lakeside Blvd., Indianapolis, IN 46278, USA
[f] U.S. Geological Survey, North Dakota Water Science Center, 821 E. Interstate Ave., Bismarck, ND 58503, USA

ARTICLE INFO

Article history:
Received 20 April 2016
Received in revised form 24 August 2016
Accepted 31 August 2016
Available online 19 September 2016
This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Purna Chandra Nayak, Associate Editor

Keywords:
Water-quality
Load estimation
Nutrients
Sediment
SPARROW

ABSTRACT

Effective management of water resources requires accurate information on the mass, or load of water-quality constituents transported from upstream watersheds to downstream receiving waters. Despite this need, no single method has been shown to consistently provide accurate load estimates among different water-quality constituents, sampling sites, and sampling regimes. We evaluate the accuracy of several load estimation methods across a broad range of sampling and environmental conditions. This analysis uses random sub-samples drawn from temporally-dense data sets of total nitrogen, total phosphorus, nitrate, and suspended-sediment concentration, and includes measurements of specific conductance which was used as a surrogate for dissolved solids concentration. Methods considered include linear interpolation and ratio estimators, regression-based methods historically employed by the U.S. Geological Survey, and newer flexible techniques including Weighted Regressions on Time, Season, and Discharge (WRTDS) and a generalized non-linear additive model. No single method is identified to have the greatest accuracy across all constituents, sites, and sampling scenarios. Most methods provide accurate estimates of specific conductance (used as a surrogate for total dissolved solids or specific major ions) and total nitrogen – lower accuracy is observed for the estimation of nitrate, total phosphorus and suspended sediment loads. Methods that allow for flexibility in the relation between concentration and flow conditions, specifically Beale's ratio estimator and WRTDS, exhibit greater estimation accuracy and lower bias. Evaluation of methods across simulated sampling scenarios indicate that (1) high-flow sampling is necessary to produce accurate load estimates, (2) extrapolation of sample data through time or across more extreme flow conditions reduces load estimate accuracy, and (3) WRTDS and methods that use a Kalman filter or smoothing to correct for departures between individual modeled and observed values benefit most from more frequent water-quality sampling.

1. Introduction

An accurate assessment of the mass, or load of water-quality constituents transported by streams and rivers is critical to the management of surface water resources both for human use and ecological health. Such information is necessary for understanding the quality of downstream receiving water bodies, the nature of upstream sources, and the relative contributions of different source areas. Accurate load estimates are needed for the cost-effective design and evaluation of water-quality management programs, which can be expensive given the large scale of many water-quality problems.

Load is expressed as the total mass passing a stream location over a given period such as a year or a decade and can be quantified by summing the product of concentration and discharge (streamflow) compiled at smaller time steps (e.g., daily) over that period. Discharge is estimated using frequent stage measurements (e.g. 15-min intervals) in conjunction with an up-to-date rating curve (calibration of the stage versus discharge relationship) which is based on discharge measurements taken several times per year. However, concentration measurements are expensive and are usually collected on a monthly or longer basis. Thus to estimate load,

[*] Corresponding author.
E-mail addresses: cjlee@usgs.gov (C.J. Lee), rhirsch@usgs.gov (R.M. Hirsch), gschwarz@usgs.gov (G.E. Schwarz), dholtschlag@usgs.gov (D.J. Holtschlag), spreston@usgs.gov (S.D. Preston), cgcrawfo@usgs.gov (C.G. Crawford), avecchia@usgs.gov (A.V. Vecchia).

http://dx.doi.org/10.1016/j.jhydrol.2016.08.059
0022-1694/Published by Elsevier B.V.

methods must be used to estimate concentration for those days when no measurement is available. This is commonly known as the load estimation problem.

A wide variety of methods have been proposed to estimate load. They can be classified according to three general types including: simple aggregation/interpolation techniques; ratio estimators that were originally published in the statistical literature for improving the results of survey sampling efforts (Cochran, 1977); and regression-based techniques that are intended to capitalize on expected patterns of covariance between concentration and discharge and/or time (Ferguson, 1986; Cohn et al., 1989). Recently, a fourth type has been proposed that are designed to be more flexible, theoretically more robust, and applicable over a broader range of conditions (Hirsch et al., 2010). These techniques are also regression-based, but are more complex and generally require longer data records than some of the more traditional regression-based methods.

Many techniques currently in use for estimating water-quality constituent load have been developed and utilized by the U.S. Geological Survey (USGS). The most commonly used USGS methods are based on multiple regression that relates observed concentrations to a set of core explanatory variables composed of contemporaneously observed daily discharge, time, and season (and possibly other variables derived from these core variables). Daily load estimates are then formed by using the regression model to predict daily values of concentration and multiplying by daily flow. Daily load estimates are then summed to form estimates of total load over periods such as a month, a year, or a decade. An extensive literature describes these methods, including: 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), Stenback et al. (2011) and Richards et al. (2012). The most commonly used USGS software package for estimating constituent load using regression is known as LOADEST (Runkel et al., 2004). For typical applications, load is estimated using 5 or 7-parameter regression models that utilize explanatory variables defined by time, discharge, and/or season. A more recent USGS software package, known as FLUXMASTER-K (similar to what is described in Schwarz et al., 2006) offers estimation methods that are similar to LOADEST, but with additional statistical enhancements designed to improve load estimates for subsequent use in modeling.

Ideally, load estimates derived using any of the available techniques would have low error associated with them including low bias (systematic error) and low variance (random error). Some random error is expected with any technique and this error can be estimated as a statistically defined standard error about the load estimate. However, systematic error can be caused by a variety of factors introduced by unexpected influences of watershed and in-stream processes. Such bias is often undetectable without near-daily data collection, which is not feasible for most sampling programs, although various diagnostic statistics and graphics can be very helpful in identifying potentially serious bias problems (see Hirsch, 2014). Thus an important quality for load estimation techniques is that they are robust with regard to bias when applied over a range of water-quality constituents, stream types, and watershed conditions.

Over recent decades many papers have described evaluations of the relative performance of different methodologies with regard to load estimation error. However, no study has shown that any one method is superior in all cases. Rather, the performance of a given estimation method often depends on the objectives of the study and the conditions in the streams being considered. More recent studies (Stenback et al., 2011; Garrett, 2012; Moyer et al., 2012; Richards et al., 2012) have renewed the interest in the topic of load estimation method performance. An important point made in all of these papers is that there are cases in which regression-based methods can be virtually unbiased, but there are also cases in which the estimated loads are substantially biased. Furthermore, these papers show that the bias can be very large (many tens of percent) and can be either positive or negative. Analysis of this potential bias problem is made difficult by the fact that there are few situations where the true long-term load is known with a high degree of accuracy.

Hirsch (2014) recently described a comparison of the bias in load estimates generated using typical applications of regression-based methods and a more complex and flexible method known as Weighted Regressions on Time Discharge and Season (WRTDS) (Hirsch et al., 2010). The regression-based methods included a 5-parameter version with independent variables based on discharge, time and season and a 7-parameter version which also included independent variables based on quadratic expressions of discharge and time. Results of the study indicated that, while the regression-based methods often produced load estimates that were nearly unbiased, they could also produce severely biased estimates under certain conditions. Those conditions include: (1) poor fit of the concentration/discharge relation; (2) changes in the concentration/discharge relation across seasons; and (3) severely heteroscedastic regression residuals. The WRTDS method was more robust to these sources of bias, but not immune to them. Verma et al. (2012) characterized the accuracy of a 7-parameter regression-based method, a version of a ratio estimator, and a flow-weighted average method (similar to a ratio estimator) and various error correction techniques based for nitrate loads in two Illinois watersheds. This study found improvements in accuracy when using ratio and the flow-weighted average method, and generally found further improvements in accuracy by the use of different methods of correcting for local deviations from sampled values. A limitation of the work by Verma et al. (2012) and Hirsch (2014) is that these studies were based on relatively few water-quality constituent records (one and two respectively), relatively few sites (2 and 5 respectively) and evaluations of only three estimation methods were evaluated.

The objectives of this paper are to extend the work of Hirsch (2014) and Verma et al. (2012) to determine the potential for load estimation bias across a broader range of methods and across a broader range of water-quality constituents, stream types and sampling regimes. We selected a set of load estimation methods for evaluation that includes examples of each of the 3 general types described above as well as the more complex and flexible methods such as WRTDS. We included a number of variations of the regression-based methods to determine if there were ways in which the large biases described by Hirsch (2014) could be reduced so as to provide versions that are more robust to the causes of those large biases. Similar to Verma et al. (2012) and Hirsch (2014), we evaluated the load estimation methods based on subsampling studies, but we obtained and utilized a larger number of water-quality records including those from a larger number of streams and water-quality constituents.

2. Methods

Depending on study objectives, load can be estimated for a range of time periods including decades, years, seasons, months, days or even specific hydrologic events. Data needs and the accuracy of estimation methods may differ among these time periods (Robertson, 2003), thus complicating comprehensive method evaluation. To limit the scope of our study, we chose to evaluate load estimation on a decadal basis. This choice was made to include the effects of constituent behaviors that could only be observed over a longer time frame, such as temporal trends that may be substantial over the period of interest, and to evaluate methods for estimating load that could subsequently be useful for calibrating water-quality models. USGS SPAtially Referenced Regression On Watershed attributes (SPARROW) models are designed to statistically simulate spatial patterns in water quality and thus require load estimates from as many locations as possible. Load estimates used in SPARROW models are typically calculated over longer time frames to attempt to factor out year-to-year hydrologic variability while continuing to focus on a specific time frame (Schwarz et al., 2006). Accurate SPARROW load estimates depend, in part, on having accurate load estimates at many sites with which to calibrate SPARROW models (Stenback et al., 2011; Richards et al., 2012); this provides an additional incentive to evaluate the ability of methods to estimate decadal load. While we expect that some of the results we find for decadal load will also apply to other time periods, we draw no conclusions in that regard.

Sub-sampling studies are performed here by randomly selecting constituent concentrations from nearly complete decadal records of daily measurements based on specific sampling strategies typically used for water-quality monitoring. These sub-samples are used in combination with complete, daily, discharge records to estimate decadal load using each method considered. Error is determined by comparing load estimates to observed decadal load determined by summing the daily products of discharge and concentration. This procedure is repeated a total of 10 times for each combination of method and sub-sampling strategy; the error for each replicate is quantified as the percentage difference between the load estimate determined from the sub-sample and the observed decadal load.

Our overall objective is to evaluate the performance of load estimation methods over a broad set of constituents, environmental conditions, and water-quality sampling scenarios. Accordingly we compile data records for a range of water-quality constituents collected at sites located across the United States (U.S.) with diverse hydrologic conditions (Fig. 1). Sub-samples are extracted from these records to evaluate method performance across the breadth of sampling amounts, frequencies, and strategies typically used by water-quality monitoring agencies.

2.1. Water-quality constituents considered

Water-quality constituents used in this study represent a range of environmental behavior patterns, particularly with respect to the concentration/discharge relations on which many sampling strategies and load estimation methods are based. Constituents chosen include specific conductance, nitrate, total nitrogen, total phosphorus, and suspended-sediment. These constituents have a number of important characteristics in terms of policy-relevance, availability of long-term, near-daily records, and a range in stream transport behaviors and concentration/discharge relations. While it would be desirable to test other constituents, such as pesticides or organic contaminants, we were not able to find daily sample records needed to compute decadal load. Censored (i.e. - "below detection") values can be present in the records of many of these constituents, but the data sets selected for this study had only a very small number (less than 0.2%) of censored values. Almost all of the estimation methods we considered have appropriate capabilities for treating censored data, but they were not used in this study. Specific conductance, suspended-sediment, and streamflow data used in this study are available from the USGS National Water Information System (http://dx.doi.org/10.5066/F7P55KJN), sources of other water-quality data are detailed below.

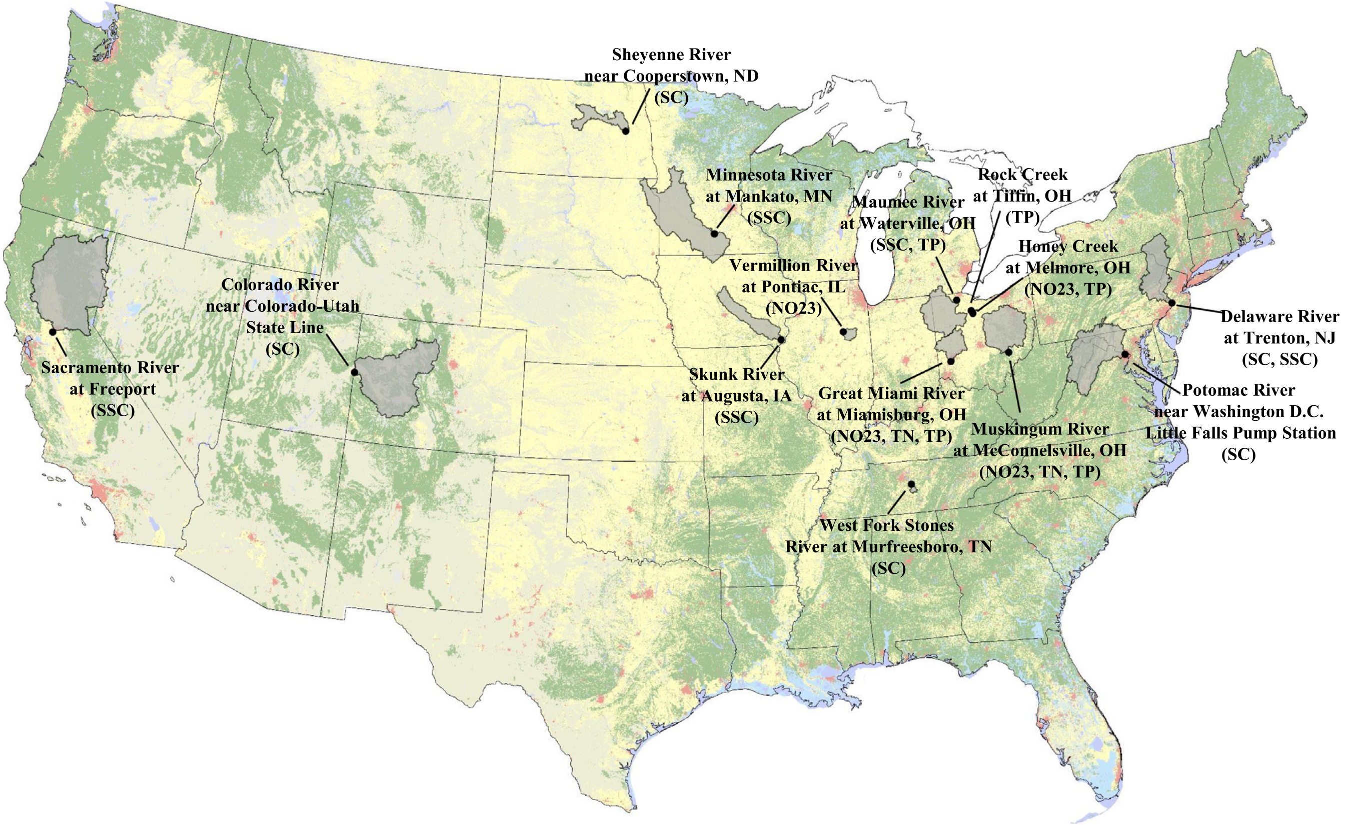


Fig. 1. Location of sites used to evaluate load estimation methods. [SC, specific conductance; SSC, suspended-sediment; NO23, nitrate; TN, total nitrogen; TP, total phosphorus.]

2.1.1. Specific conductance

Specific conductance was chosen as a surrogate for dissolved solids concentration because the two are usually closely correlated (Hem, 1985) and because detailed specific conductance records are often readily available due to the deployment of continuous monitors by agencies across the U.S. Specific conductance is often is inversely correlated with streamflow conditions (O'Connor, 1976). For the purposes of this study, specific conductance values reported in micro-siemens/cm are used as if they are concentrations of total dissolved solids for evaluating the performance of load estimation methods.

2.1.2. Suspended sediment

Mean daily suspended-sediment concentration data have been reported by the USGS for decades (Lee and Glysson, 2013). Suspended-sediment is transported almost exclusively during high-flow conditions (Wolman and Miller, 1960), and substantial variation in concentrations among high flows often makes it difficult to produce unbiased, robust load computations (Walling, 1977; Horowitz, 2003). Thus, in direct contrast to specific conductance and total dissolved solids, suspended-sediment represents those water-quality constituents that increase in concentration with discharge due to storm runoff.

2.1.3. Nutrients

Nitrate, total phosphorus, and total nitrogen were chosen for this study because of their importance to streams and estuaries, but also because uncensored, near-daily records have been collected for these species for more than 10 years by Heidelberg University (2005) as well as for nitrate on the Vermillion River near Pontiac, IL (G. McIsaac, written commun., 2013). Each of these three constituents exhibits different transport properties that can result in challenges for load estimation. Nitrate moves in soluble form through soils and groundwater and, depending on the sources in a watershed (agriculture, urban runoff), can be transported primarily through either runoff or groundwater pathways (Tesoriero et al., 2013), resulting in some unique challenges for load estimation as noted by Stenback et al. (2011) and Hirsch and De Cicco (2014). Phosphorus is generally transported in streams while sorbed to sediment particles and thus behaves similarly to suspended-sediment, with more potential for seasonal influences related to fertilizer application. Total nitrogen is comprised of both dissolved nitrate and particulate forms and thus is transported via both of the above pathways.

We conducted a search to identify sites with long-term, daily concentration records representing a diverse set of hydrologic conditions across the U.S. Although complete, decadal records are generally available for suspended-sediment; other constituents generally have periodic gaps in daily records. In order to evaluate methods across as many, and as complete, decadal records as possible, we limit sites to those with uncensored data available for more than 80% of days and more than 80% of the total flow volume sampled for each year of the 10-year period. For these records, decadal load is computed as the sum of daily load on all sampled days and any daily values that were missing or censored were omitted (and are similarly omitted from the estimates that are summed in the sub-sampling experiments). Sites are selected across as many landscapes and stream sizes as possible, but selection is limited by available data. This is especially true for nutrients, for which limited data availability constrained site locations to the upper Midwest. In general we try to include records from 5 sites for each constituent, but are limited to less than 5 for nitrate and for total nitrogen (Fig. 1 and Table 1).

It is important to note that records used to represent observed decadal loads in this study are themselves subject to multiple sources of error. Potential errors include, but are not limited to, sensor calibration and fouling errors (in the case of specific conductance), non-representative sampling techniques, and laboratory errors. These types of errors are inherent in water-quality records derived from sampling programs and are not evaluated as part of this study. We choose instead to focus on the uncertainty derived from the application of load estimation methods to typical sampling records and assume that the complete decadal record of daily samples provides the closest possible measure of a "true" load.

2.2. Estimation methods

We consider 11 load estimation methods that are either frequently used in practice or are representative of a given approach. The 11 methods range from simple to computationally complex, and include examples from each of the four types of load methods described in the introduction. A summary of all 11 methods is provided below; more detailed descriptions of estimation methodologies can be found in the references or in the supplemental information section of this paper.

2.2.1. Simple data-driven methods

2.2.1.1. Linear interpolation among sampled values (INTERP). The concentration on each non-sampled day is estimated by linear interpolation between the concentration values collected on adjacent sampled days. Daily load is then computed as the product of the daily concentration and the daily discharge, and decadal load is estimated as the sum of the daily estimates. For non-sampled days prior to the first sample day, the estimate of concentration is the value observed on the first sample day, and for days following the last sample day, the estimated concentration is value observed on the last sample day.

2.2.1.2. Beale Ratio Estimator (RATIO). Beale's ratio estimator is implemented in stratified form as described in Cochran (1977). Load values calculated on sample days are initially assigned to one of eight strata formed by two flow classes in each of four seasons, the flow classes being delineated by the 80th percentile of flow for each individual water year. If the number of samples in a flow class, across all seasons, is less than 10 then the two flow classes within each season are collapsed into a single class to give a strictly seasonal stratification. Further strata collapse is based on combining seasons within the remaining flow class or classes. This is accomplished by identifying the season with the smallest number of samples. If this season has less than 10 samples, it is combined with the neighboring season having the fewest samples; otherwise, no further collapse of strata is required. This process is repeated until all remaining strata include at least 10 samples.

The Beale estimator for the ratio of a given stratum is given by

$$\widehat{R}\left(\frac{1+\frac{1-f}{n}c_{LQ}}{1+\frac{1-f}{n}c_{QQ}}\right), \tag{1}$$

where $\widehat{R} = \bar{l}/\bar{q}$ is the ratio of the stratum sample means of load, $\bar{l}$, and flow, $\bar{q}$; f = n/N – the ratio of the number of samples in the stratum, n, to the total number of days (sampled or unsampled) in the prediction period occurring in the stratum, N; $c_{LQ} = s_{LQ}/(\bar{l}\bar{q})$, the ratio of the stratum sample covariance between load and flow, s_{LQ}, to the product of the stratum sample means of load and flow, and $c_{QQ} = s_Q^2/\bar{q}^2$ is the ratio of the stratum sample variance of flow to the square of the stratum sample mean of flow. The estimate of load for all days within a given stratum is given by the sum of daily load in the sample plus the product of the estimated Beale ratio for the stratum, multiplied by the total flow for all unsampled days in the stratum. The summation of these estimates across all strata

Table 1
Sites and water quality constituents used for load evaluation.

Site name	Site abbreviation (used in Figs. 1 and 7)	USGS site identifier	Contributing drainage area (mi^2)	Period of record	Percentage agriculture[b]	Percentage forest[b]	Percentage urban[b]	Coefficient of variation of daily streamflow (in percent)	Coefficient of variation of daily load (in percent)	Number of missing or censored days	Total load
Specific conductance (total load is in us/cm ∗ cubic feet per second)											
Delaware River at Trenton, NJ	DELA	01463500	6,780	1982–1991	15	67	10	97	63	149	39,802,800
Potomac River near Washington D.C. Little Falls Pump Station	POTO	01646500	11,560	2003–2012	30	59	10	129	97	41	46,858,700
West Fork Stones River at Murfreesboro, TN	WEST	03428200	177	2003–2012	41	31	22	243	163	29	960,700
Sheyenne River near Cooperstown, ND	SHEY	05057000	1,270	2003–2012	67	0	4	206	151	187	1,098,200
Colorado River near Colorado-Utah State Line	COLO	09163500	17,849	2003–2012	4	54	1	101	37	198	19,945,600
Suspended-sediment (total load is in kilograms)											
Delaware River at Trenton, NJ	DELA	01463500	6,780	1972–1981	15	67	10	92	541	46	49,958,800
Maumee River at Waterville, OH	MAUM	04193500	6,330	1975–1984	79	6	11	177	348	0	21,100,800
Minnesota River at Mankato, MN	MINN	05325000	14,900	1974–1983	79	2	6	142	197	0	12,308,500
Skunk River at Augusta, IA	SKUN	05474000	4,312	1978–1987	77	7	7	126	275	0	12,472,000
Sacramento River at Freeport, CA	SACR	11447650	27,233	1972–1981	12	44	4	76	199	0	81,474,100
Nitrate plus nitrite (total load is in kilograms)											
Muskingum River at McConnelsville, OH	MUSK	03150000	7,422	2003–2012	41	43	12	94	104	159	32,302,000
Great Miami River at Miamisburg, OH[a]	GRMI	03271601	2,715	2003–2012	72	9	17	137	138	193	12,592,500
Honey Creek at Melmore, OH	HONE	04197100	774	1986–1995	82	10	7	212	217	141	434,600
Vermillion River at Pontiac, IL	VERM	05554500	579	1989–1998	92	1	6	183	181	1	1,954,300
Total nitrogen (total load is in kilograms)											
Muskingum River at McConnelsville, OH	MUSK	03150000	7,422	2003–2012	41	43	12	94	106	223	31,724,900
Great Miami River at Miamisburg, OH[a]	GRMI	03271601	2,715	2003–2012	72	9	17	137	145	248	12,500,600
Honey Creek at Melmore, OH	HONE	04197100	774	1986–1995	82	10	7	212	220	152	8,131,800
Total phosphorus (total load is in kilograms)											
Muskingum River at McConnelsville, OH	MUSK	03150000	7,422	2003–2012	41	43	12	94	143	140	32,385,600
Great Miami River at Miamisburg, OH[a]	GRMI	03271601	2,715	2003–2012	72	9	17	137	196	194	12,591,700
Maumee River at Waterville, OH[a]	MAUM	04193500	6,330	1991–2000	79	6	11	155	300	210	19,117,000
Honey Creek at Melmore, OH	HONE	04197100	774	1986–1995	82	10	7	212	340	154	406,100
Rock Creek at Tiffin, OH	ROCK	04197170	35	1994–2003	79	11	9	344	578	109	97,900

[a] To incorporate as many nutrient species as possible, only 60% of flows were sampled at the Great Miami River in 2005, and only 77% of flows were sampled on the Maumee River in 1992.
[b] Based on 2006 National Land Cover Database and aggregated from Falcone (2011) and the USGS Sediment Portal (Lee and Glysson, 2013).

gives the ratio method total load estimate for all days in the prediction period. The Beale ratio estimator approach assumes a positive correlation between flux and flow and has been utilized extensively for flux estimation in the Great Lakes region and in other parts of the U.S. (Richards and Holloway, 1987) generally employing more complex strata definition strategies than those used in this study.

2.2.2. Regression methods

Regression methods use standard least squares or maximum likelihood approaches to relate infrequently available concentration data to various predictor variables derived from daily flow estimates and decimal time. Regression methods assume that model residuals are normally distributed with a constant variance (Runkel et al., 2004). Alternative specifications of the regression models are considered, with different assumptions regarding the explanatory variables used to explain concentration and the correlation structure of the residuals. The basic form consists of the 7-parameter model described by Cohn et al. (1989) in which logarithm-transformed daily concentration is related to second-order polynomials of logarithm-transformed daily flow, decimal time, and seasonal factors derived from transformations of decimal time. The 7-parameter model is defined as

$$\ln(C_t) = \beta_1 + \beta_2 \ln Q_t + \beta_3 \ln{(Q_t)}^2 + \beta_4 T_t + \beta_5 T_t^2 + \beta_6 \sin(2\pi T_t) + \beta_7 \cos(2\pi T_t) + e_t, \tag{2}$$

where $\ln(C_t)$ is the natural logarithm of constituent concentration for period t, assumed to be a day; $\ln(Q_t)$ is the natural logarithm of mean daily discharge; T_t is decimal time, in years; e_t is a model residual; and β_k, $k = 1, \ldots, 7$, are model parameters to be estimated.

Regression methods are implemented through the LOADEST or FLUXMASTER load-estimation software packages, both developed by the USGS. LOADEST is designed to estimate water-quality constituent flux in streams and rivers using either Adjusted Maximum Likelihood Estimation (AMLE) or Maximum Likelihood Estimation (MLE) methods which produce identical results in the absence of censored data (Runkel et al., 2004). A minimum variance unbiased estimate (MVUE) of instantaneous flux is used to correct for potential retransformation bias (Cohn et al., 1989).

FLUXMASTER was developed specifically with the objective of estimating detrended stream load for subsequent use in USGS SPARROW models (Schwarz et al., 2006). Previous uses of FLUXMASTER for SPARROW (see, for example, Saad et al., 2011) employ a variant of the algorithm that in the absence of censored observations is identical to LOADEST in the estimation of the concentration model described by Eq. (1), but differ from LOADEST in that predicted loads are adjusted to remove the effects of trend in flow and concentration. The version of FLUXMASTER evaluated in the present study has four significant differences not previously employed in SPARROW applications: an allowance for first-order serial correlation of the residuals is included in the estimation of the concentration model, a serial correlation structure is used to apply a Kalman smoothing method in load prediction, retransformation bias is corrected using a parametric bootstrap method, and the detrending feature is not implemented to enable a direct comparison between actual and estimated loads. Because the present study does not include any censored data, the method used to estimate the concentration model is standard maximum likelihood (see the supplemental information for a detailed description of the method; censored data require a simulation-based version of maximum likelihood), with the coefficient estimates subsequently adjusted using a procedure analogous to that implemented in LOADEST (Cohn, 2005) to remove first-order bias in the maximum likelihood estimates. Further details are presented in the supplemental information.

2.2.2.1. LOADEST 5 parameter model (L5). The 5-parameter version of the regression model represents a simpler specification than the 7-parameter version which could avoid potential over-interpretation of the identified relations between concentration and discharge or time. The specification excludes quadratic terms from Eq. (1), the terms most sensitive to over-interpretation when applying the model to conditions outside the water-quality sample. The model takes the form

$$\ln(C_t) = \beta_1 + \beta_2 \ln Q_t + \beta_3 T_t + \beta_4 \sin(2\pi T_t) + \beta_5 \cos(2\pi T_t) + e_t, \tag{3}$$

The 5-parameter model is implemented using the LOADEST algorithm and software.

2.2.2.2. FLUXMASTER-K 5 parameter model (F5). Similar to L5 (described above in Section 2.2.2.1), the FLUXMASTER-K version of the 5-parameter model is included in the analysis to test for potential differences between the 2 packages and specifically for benefits provided by the Kalman smoothing algorithm and an alternative approach to retransformation bias correction.

2.2.2.3. LOADEST cubic model (LCUBE). This method accounts for greater complexity in the concentration/discharge relation by augmenting the 5-parameter model (Eq. (2)) with two additional explanatory variables given by the quadratic and cubic transforms of the logarithm of daily discharge.

2.2.2.4. LOADEST 7-parameter model (L7). This version is the basic form of the 7-parameter models as described above in Eq. (1), implemented using the LOADEST software.

2.2.2.5. LOADEST 7-parameter model with composite method (L7COMP). This method adjusts L7 results for residual departures by interpolation among residual departures (in arithmetic space) using methods described in Aulenbach and Hooper (2006).

2.2.2.6. FLUXMASTER-K 7-parameter model (F7). This method is the same as that used for F5, with the additional quadratic forms of the flow and time variables present in L7 included in the model specification.

2.2.2.7. LOADEST model in which explanatory variables are selected to minimize the Akaike Information Criteria (AIC; Akaike, 1974) (LAIC). This version of the regression model is based on the idea of limiting the explanatory variables only to those that account for a significant amount of the variability in the dependent variable (load). In its most complex form the model is based on the 7-parameter model. All possible iterations of the explanatory variables within the 7-parameter model are considered, and the model with the smallest AIC value is selected. This option is available as part of the LOADEST software package.

2.2.3. Flexible functional form parametric methods

2.2.3.1. Weighted regressions on time, season, and discharge (WRTDS). Weighted regression on time, discharge, and season (WRTDS) is implemented through the R package Exploration and Graphics for RivEr Trends (EGRET) (Hirsch and De Cicco, 2014). WRTDS is used to develop nonlinear, time-varying relations between the logarithm of concentration and the explanatory variables consisting of decimal time, the logarithm of daily discharge, and sine and cosine transformations of decimal time (Hirsch et al., 2010). The method derives these flexible relations using a unique weighted regression for each day of the estimation period. Weights for each day in the sample are based on differences in the values of the explanatory variables between the prediction and sample day. The method employs a bias correction factor specific to each year, day, and discharge to adjust for retransformation bias (see Moyer et al., 2012; Hirsch and De Cicco, 2014).

2.2.3.2. Generalized Additive Multiple Modeling with Kalman Smoothing (GAMMKS). The Generalized Additive Multiple Modeling with Kalman Smoothing (GAMMKS) estimates constituent load as a function of streamflow, season, and time through a weighted average of a Generalized Additive Model (GAM) and Multiple Linear Regression (MLR) model. The GAM model predicts load through a linear combination of smoothing functions and daily data on flow, season, and time; a 1 to 30-day anomaly (Ryberg and Vecchia, 2012) is included when there are more than 100 observations. The MLR takes a form identical to the L7 model except it does not include the time-squared term. Weights assigned to GAM and MLR models are inversely proportional to each model's variance for a particular observation. Adjustment for log-retransformation bias is performed as in LOADEST, and as with the FLUXMASTER-K method; a Kalman smoothing procedure is used to adjust for local departures from observed data. More details on this method are provided in the supplemental information section of this paper.

2.3. Evaluation of load estimation methods

The performance of estimation methods is determined by sampling from daily water-quality records under prescribed sampling scenarios. For each sampling scenario, data for sampled days are used to estimate decadal load and are then compared to the estimated value of the decadal load obtained by summing the actual record of daily loads. Each sampling scenario is run 10 times by randomly sampling different days of the observed water-quality record under the prescribed sampling scenario.

Differences in the availability of sample data and sample collection strategies undertaken by different agencies can affect the bias and variability of load estimates (Robertson, 2003). Because models like SPARROW rely on data collected by many organizations with different sampling objectives, it is necessary to characterize the performance of load estimation methods with respect to different sampling over different time spans, frequencies, and strategies. A sample generation program is developed to subsample daily to mimic the types of discrete datasets produced by various sample collection agencies, with the number of samples strictly fixed across repetitions and with each subsampling repetition having an approximately equal selection probability. The sample generation program is summarized below and more detailed information is provided as supplemental information.

2.3.1. Sampling strategy

Sampling strategies are common approaches monitoring agencies use to schedule water-quality sample collection. Four different sampling strategies are selected based on those typically encountered in the development of SPARROW models.

- Uniform: Sample days are approximately uniformly spaced in time at a specified frequency.
- High flow: 30% of sample days are approximately uniformly spaced in time, the remaining 70% of samples are randomly selected days above the 80th percentile of daily flows for a given year.
- Seasonally-weighted: Each season is assigned a weight dependent upon the fraction of flow that occurred during that season over the decadal water-quality record. The number of days sampled each season is the fraction of flow for that season multiplied by the specified sampling frequency. Within each season samples are approximately uniformly spaced in time.
- Low flow: No days above the 80th percentile of flow are sampled, and sample days are approximately uniformly spaced in time among the remaining days.

2.3.2. Sampling furloughs

The length of water-quality records available at a particular sampling site is often dictated by budget constraints. Changes to funding may cause sample collection to begin or end at various points in a record or may result in temporal gaps within a water-quality record. To test the effect of what are termed "sampling furloughs", four scenarios are tested over the 10-year monitoring period.

- No furlough: The entire 10-year site record is used.
- Middle furlough: Years 1–3 and 8–10 are used.
- Tail furlough: Years 3–8 are used.
- End furlough: Years 1–6 are used.

2.3.3. Sampling frequency

Monitoring agencies collect data at different frequencies depending on objectives and availability of funds. Sampling frequencies of 6, 12, 24, and 52 per year are evaluated in this study.

Ten repetitions are run for all combinations of four sampling strategies, four sampling furloughs, and four sampling frequencies. A total of 640 decadal load estimates are computed for each of the 22 water-quality records, resulting in a total of 14,080 runs for each estimation method.

3. Results

We assess the performance of methods for estimating decadal loads by quantifying the mean percent error (MPE) as a measure of bias and the root mean squared percentage error (RMSPE, in percent) as a measure of overall error. MPE is defined as

$$\mathrm{MPE} = \frac{100}{mn} * \sum_{j=1}^{m}\sum_{i=1}^{n}\frac{EST_{ij} - OBS_j}{OBS_j}, \tag{4}$$

where EST_{ij} is the estimated load for case i (i.e. the estimate for a particular sampling strategy, furlough, frequency, and repetition) and station j, OBS_j is the observed load for station j, and n is the total number of days in the prediction period occurring in the stratum.

RMSPE is defined as

$$\mathrm{RMSPE} = 100 * \sqrt{\frac{\sum_{j=1}^{m}\sum_{i=1}^{n}\left(\frac{EST_{ij}-OBS_j}{OBS_j}\right)^2}{mn}}. \tag{5}$$

MPE and RMSPE are presented for each method and water-quality constituent in the form of "level plots" in which the magnitude and direction of error is indicated by shading and by color (Figs. 2–4, 6 and 7). For MPE, the level plot shading indicates the magnitude of the aggregate error and the color indicates direction (light to dark blue for over-estimation and yellow to red for under-estimation). For RMSPE, the magnitude of the aggregate error is indicated by grey shading with darker indicating greater error. Initially we present a fully aggregated plot (combination of all sampling strategies, furloughs, frequencies, and sites) to describe general results for each constituent and estimation method (Fig. 2). We then present level plots that are designed to show differences in method performance due to specific monitoring record characteristics. For example, Fig. 3 shows differences in errors among sampling strategy based on errors aggregated over replicate, frequency, and site location. Similar patterns among the errors due to sampling strategy are observed across frequencies and site locations and thus little information was lost by

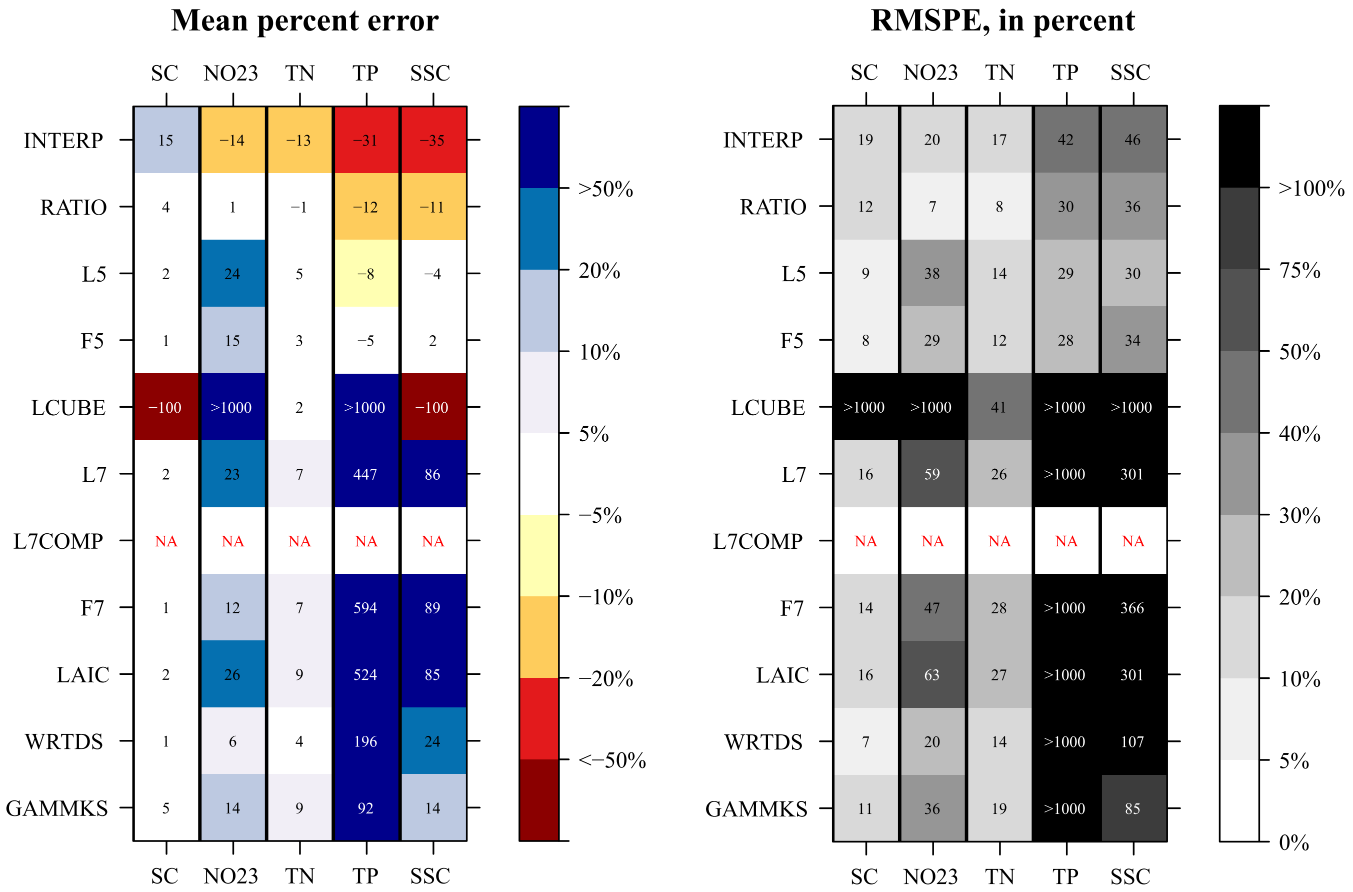


Fig. 2. Comparison of mean percent error and root mean squared percent error for estimation methods among constituents relative to observed decadal loads. [SC, Specific conductance; NO23, Nitrate; TN, Total Nitrogen; TP, Total Phosphorus; SSC, Suspended sediment.]

aggregating across those factors. Similar aggregated plots are described for other effects including furlough, sampling frequency, and site location (Figs. 4, 6 and 7).

Fig. 2 provides a general summary of method performance by constituent based on aggregated error across all of the cases considered. Among all cases, method performance generally differs among the constituents tested due to the environmental behavior of that constituent and its typical relation with discharge. In general, RMSPE's and absolute values of MPE are lowest for specific conductance which often has an inverse, less variable relation with discharge than other constituents. In contrast, RMSPE's and absolute values of MPE are often greatest for suspended sediment concentration which tends to have a positive, but highly variable relation with discharge. Error levels for the other constituents tend to increase in the order of total nitrogen, nitrate, and total phosphorus, an ordering which is also consistent with the increasing influence of periodic high-discharges on water-quality loads.

Fig. 2 provides the broadest overview of the results and is useful in that regard, but it includes a number of limitations. First, for some of the methods, the aggregations cannot be developed over a full set of cases due to specific requirements of the methods. For example, composite method (L7COMP) estimates are not displayed because this method is not designed for records where furloughs are present and cannot be applied for those cases. Similarly, although interpolation (INTERP) results in Fig. 2 include furloughed sampling scenarios, this method is generally not used when prolonged sampling gaps occur, thus performance summarized in Fig. 2 is generally not representative for that method. A second general limitation of Fig. 2 is that it aggregates results over many different cases, some of which tend to dominate the final numbers. For example, results for the LOADEST cubic model (LCUBE) tend to be dominated by the error levels for specific cases in the low-flow sampling scenario for which its performance is particularly poor. Similar effects are observed for other regression methods and for the more complex methods (WRTDS and GAMMKS). Because of the limitations described above, we develop similar figures that provide more detail on the key factors that affect method performance (Figs. 3, 4, 6 and 7).

3.1. Effect of sampling strategy

Method performance with respect to uniform (U), high flow (H), seasonal (S), and low-flow (L) sampling strategies is summarized in Fig. 3. MPE and RMSPE values presented in Fig. 3 are aggregated over sampling frequencies and monitoring sites because similar patterns in method performance among strategies are observed over those two effects. However, only the no-furlough case is included because sample records with furloughs frequently have high error rates that tend to dominate the aggregate values. Results related to furloughs are presented separately in the next section in order to isolate those effects. Results for the composite method (L7COMP) are not reported in Fig. 3 for TP and SSC under the low-flow sampling strategy because the model code did not converge to a solution for several of these records.

Load estimates calculated using samples collected under the low-flow sampling strategy are nearly always less accurate than samples collected under uniform, seasonal, or high-flow sampling strategies. The absolute values of MPE for low-flow estimates are greater than other sampling strategies for 50 of the 53 constituent/method pairs, and are more than double uniform,

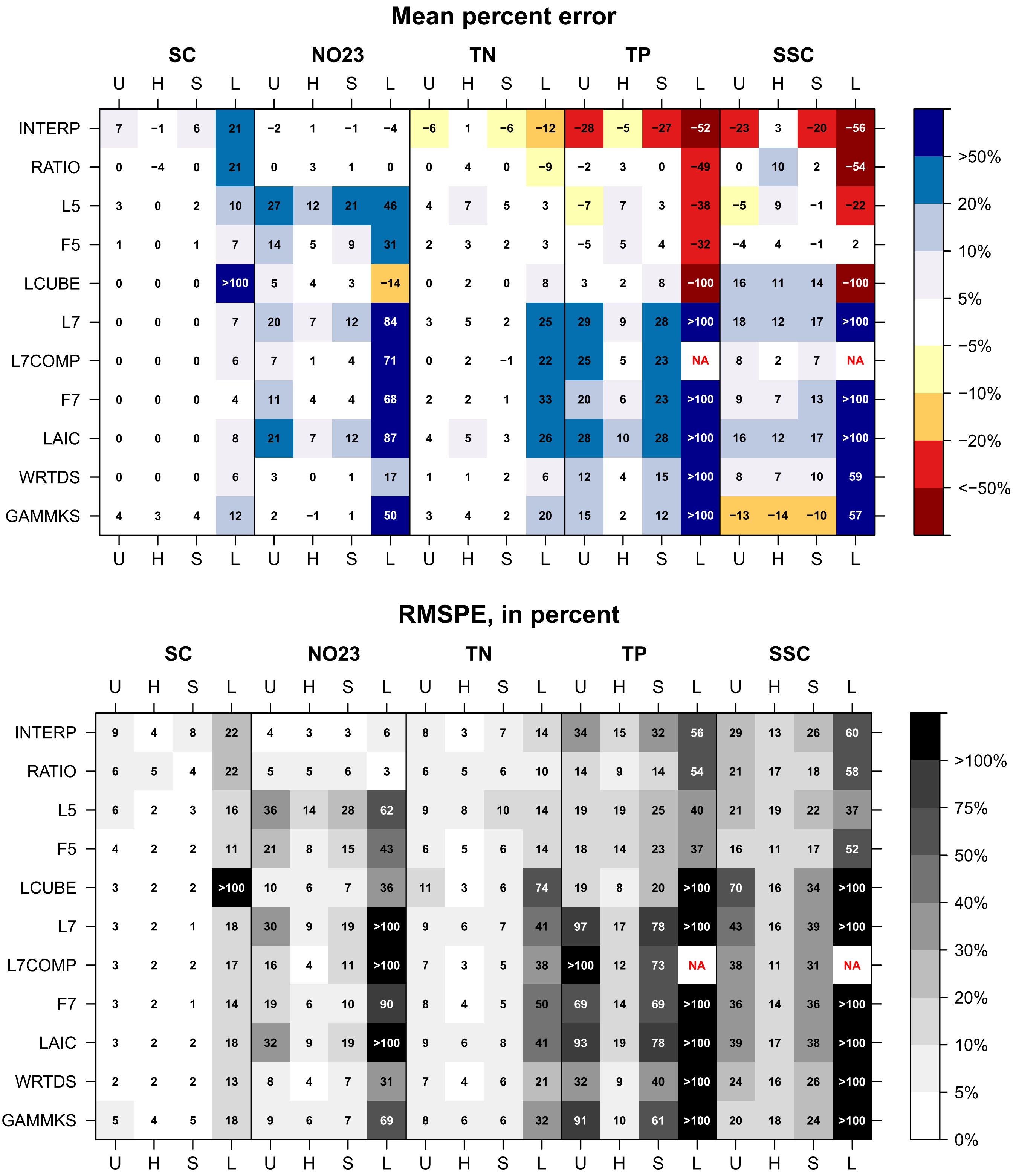
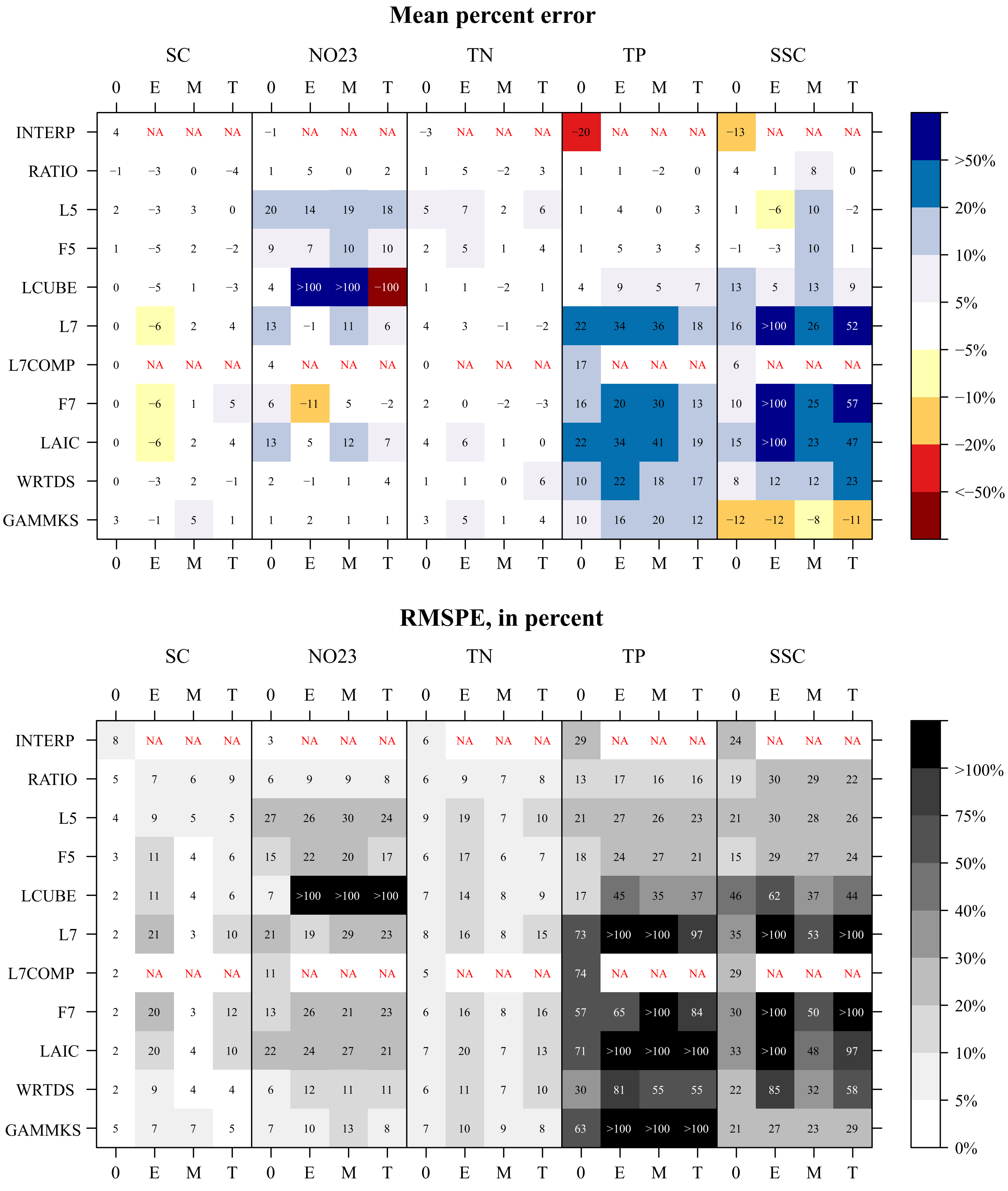


Fig. 3. Comparison of mean percent error and root mean squared percent error for estimation methods among constituents and sampling strategies relative to observed decadal loads. [SC, Specific conductance; NO23, Nitrate; TN, Total Nitrogen; TP, Total Phosphorus; SSC, Suspended sediment; U, Uniform Sampling, H, High flow sampling; S, Seasonal sampling; L, Low-flow sampling; excludes load estimates from furloughed sampling strategies.]

seasonal, or high-flow estimates in all but two of those cases. For TP and SSC, constituents with strongly positive concentration/discharge relations, simpler methods INTERP, RATIO, L5, and F5 tend to underestimate load under the low-flow sampling strategy, resulting in negative MPE values. Conversely, for TP and SSC, the more complex regression methods tend to overestimate load, resulting in positive MPE's which in some cases are quite large. This overestimation reflects the potential for the prediction of very large concentration values when extrapolating a complex concentration/discharge relation such as a quadratic relation to high



Mean percent error

	SC				NO23				TN				TP				SSC			
	0	E	M	T	0	E	M	T	0	E	M	T	0	E	M	T	0	E	M	T
INTERP	4	NA	NA	NA	−1	NA	NA	NA	−3	NA	NA	NA	−20	NA	NA	NA	−13	NA	NA	NA
RATIO	−1	−3	0	−4	1	5	0	2	1	5	−2	3	1	1	−2	0	4	1	8	0
L5	2	−3	3	0	20	14	19	18	5	7	2	6	1	4	0	3	1	−6	10	−2
F5	1	−5	2	−2	9	7	10	10	2	5	1	4	1	5	3	5	−1	−3	10	1
LCUBE	0	−5	1	−3	4	>100	>100	−100	1	1	−2	1	4	9	5	7	13	5	13	9
L7	0	−6	2	4	13	−1	11	6	4	3	−1	−2	22	34	36	18	16	>100	26	52
L7COMP	0	NA	NA	NA	4	NA	NA	NA	0	NA	NA	NA	17	NA	NA	NA	6	NA	NA	NA
F7	0	−6	1	5	6	−11	5	−2	2	0	−2	−3	16	20	30	13	10	>100	25	57
LAIC	0	−6	2	4	13	5	12	7	4	6	1	0	22	34	41	19	15	>100	23	47
WRTDS	0	−3	2	−1	2	−1	1	4	1	1	0	6	10	22	18	17	8	12	12	23
GAMMKS	3	−1	5	1	1	2	1	1	3	5	1	4	10	16	20	12	−12	−12	−8	−11

RMSPE, in percent

	SC				NO23				TN				TP				SSC			
	0	E	M	T	0	E	M	T	0	E	M	T	0	E	M	T	0	E	M	T
INTERP	8	NA	NA	NA	3	NA	NA	NA	6	NA	NA	NA	29	NA	NA	NA	24	NA	NA	NA
RATIO	5	7	6	9	6	9	9	8	6	9	7	8	13	17	16	16	19	30	29	22
L5	4	9	5	5	27	26	30	24	9	19	7	10	21	27	26	23	21	30	28	26
F5	3	11	4	6	15	22	20	17	6	17	6	7	18	24	27	21	15	29	27	24
LCUBE	2	11	4	6	7	>100	>100	>100	7	14	8	9	17	45	35	37	46	62	37	44
L7	2	21	3	10	21	19	29	23	8	16	8	15	73	>100	>100	97	35	>100	53	>100
L7COMP	2	NA	NA	NA	11	NA	NA	NA	5	NA	NA	NA	74	NA	NA	NA	29	NA	NA	NA
F7	2	20	3	12	13	26	21	23	6	16	8	16	57	65	>100	84	30	>100	50	>100
LAIC	2	20	4	10	22	24	27	21	7	20	7	13	71	>100	>100	>100	33	>100	48	97
WRTDS	2	9	4	4	6	12	11	11	6	11	7	10	30	81	55	55	22	85	32	58
GAMMKS	5	7	7	5	7	10	13	8	7	10	9	8	63	>100	>100	>100	21	27	23	29

Fig. 4. Comparison of mean percent error and root mean squared percent errors for estimation methods among constituents and sampling furloughs relative to observed decadal loads. [SC, Specific conductance; NO23, Nitrate; TN, Total Nitrogen; TP, Total Phosphorus; SSC, Suspended sediment; 0, No furlough, E, End furlough; M, Mid furlough; T, Tail furlough; excludes load estimates from the low-flow sampling strategy.]

discharges. RMSPEs are greatest for low-flow sampling in 52 of the 53 constituent/method pairs, and are more than double other strategies in 46 of these cases. All of the other strategies allow for some amount of high-flow sampling; superior performance for all methods in these cases clearly indicate the value of high-flow sampling when load estimation is among the goals of a water-quality sampling program. These results imply that a sampling program in which high-flow samples are not collected will cause high error levels no matter which load estimation method is used. For that reason the low-flow sampling strategy is omitted from further method comparisons.

Error levels among uniform, seasonal, and high-flow sampling strategies vary the most among NO23, TP, and SSC estimates. For these constituents, the high-flow sampling strategy produces the least-biased loads for 8 of 11 methods for NO23, 8 of 11 methods for TP, and 7 of 11 methods for SSC. High-flow sampling also results in the smallest RMSPE values for all methods used for NO23, TP and SSC. High-flow and seasonal sampling improve the performance of some methods over uniform sampling for SC and TN, although most method estimates for these constituents are within 5 percent of observed decadal loads regardless of strategy. Although results indicate that sampling plans benefit from an emphasis on high-flow sampling, it is important to note that the error level observed for a given strategy can vary among sampling sites depending on the observed hydrologic behavior, and that an adequate historic record of discharge must be available in order to design an appropriate high-flow sampling strategy at a given site.

3.2. Effect of sampling furlough

Method performance with respect to the presence and type of furlough is presented in Fig. 4 in which MPE and RMSPE values are summarized by method, water-quality constituent, and furlough type including: no furlough (0), end furlough (E), mid furlough (M), and tail furlough (T). MPE and RMSPE values in Fig. 4 are aggregated over sampling frequencies, sampling strategies (excluding low-flow), and monitoring sites. Furlough records sampled with the low-flow sampling strategy are not included in the aggregates for the reasons stated above. Further, furloughed records are not presented for the composite (L7COMP) and interpolation (INTERP) methods in Fig. 4 because they are not designed to provide estimates over long periods without sample collection.

Furloughs do not always increase the bias (MPE) of load estimates but generally increase variability (RMSPEs). Non-furloughed estimates have the smallest absolute value of MPE for 6 of 9 methods for SC, 2 of 9 for NO23, 2 of 9 for TN, 4 of 9 for TP, and 6 of 9 for SSC. RMSPEs are smallest for non-furloughed cases for 9 of 9 SC methods, 6 of 9 for NO23, 7 of 9 for TN, 9 of 9 for TP, and 8 of 9 for SSC. In the case of LCUBE, extreme misspecification of the relation among flow and load for one or two NO23 cases for each furlough scenario wildly inflated (or deflated) decadal load estimates relative to observed values. Among the potential furlough sampling scenarios, the mid-furlough is the only one for which data are available before and after the omitted period and most load estimation methods essentially interpolate across the furlough period. Potentially for that reason, the mid-furlough case has the smallest RMSPEs for most methods used to estimate SC, TN, and SSC loads. End furloughs have the longest continuous period (4 years) in which methods had to extrapolate loads through time, which may be why load estimates for this case have the largest RMSPEs for the majority of methods among all constituents. Among specific methods, RATIO estimates generally are the least biased (absolute value of MPE <9) among all furloughs and constituents. However, it is important to note that differences in method performance are not consistent among furloughs making it difficult to define set patterns in method performance. Factors such as the presence and degree of trends, year to year variability in concentrations and loads, and the representativeness of observed flow conditions (especially for specific conductance sites, see additional analysis in the supplemental information) likely influence load estimates when furloughs are present in the sample record. In general furloughs and associated extrapolation through time should be avoided when possible and load estimation using such records, if necessary, should be performed with caution. Given the unique nature of furlough records and the need to focus on error specifically due to other record characteristics, we exclude furlough results from further method evaluations.

3.2.1. Method performance with no sampling furloughs

Fig. 4 provides MPE and RMSPE summaries for the "non-furlough" cases separately and provides an opportunity to evaluate method performance excluding the low-flow sampling strategy and furloughs, both of which can have significant effects on aggregated error measures. Thus we describe here the error levels methods based on the "non-furlough" cases to provide additional detail on relative performance and potential causes of poor performance.

For the non-furlough cases, method performance is highly dependent on the constituent being considered as described above for Fig. 2. The MPE and RMSPE levels for the SC and TN estimates are low for all methods in the non-furlough case, never exceeding MPE values above 5% and never exceeding RMSPE values above 9%. Thus, all methods perform well for estimating TN and SC load and there is little basis for distinguishing the performance of the methods. For the other constituents considered (NO23, TP, SSC) performance is much more variable and there are substantial differences among the methods. For NO23, INTERP, RATIO, LCUBE, WRTDS, and GAMMKS estimates have low bias (abs(MPE) < 5%) and low RMSPEs (<7%). In contrast, many of the regression methods and especially LOADEST (L5, L7, and LAIC) have the largest MPE's and RMSPE's for NO23.

To better understand the causes of poor performance of the regression methods we evaluated a number of individual cases. These evaluations indicate that simpler regression-based methods have the potential to misrepresent relations among observed load and flow conditions. To illustrate, we provide examples based on NO23 data collected from the Vermillion River at Pontiac, IL (Fig. 5A) and SSC data at the Skunk River at Augusta, IA (Fig. 5B). The grey points show the observed concentration/discharge relation at the two sites, and the various lines show loess fits of the ten L5, L7, and WRTDS estimates conducted for the uniform sampling, no furlough, 52 sample per year scenario. At the Vermillion River site, approximately 60 percent of the decadal NO23 load is transported by flows above 1000 cubic feet per second which is a discharge exceeded approximately only 10 percent of the time. Positive bias observed for L5 and L7 parameter estimates at these higher flows result in decadal load estimates that are 59 and 54 percent (respectively) greater estimated than observed decadal load NO23 load estimates for the non-furloughed, uniform sampling, 52 samples per year case. Lack of fit of the concentration/discharge relations is one of the main reasons identified by Hirsch (2014) for bias in regression-based NO23 load estimates and these results are consistent with that conclusion. In contrast to the standard regression models, the flexibility inherent in RATIO, WRTDS, and GAMMKS methods across time and flow condition substantially improve NO23 load estimates at this site and sampling scenario (estimated load −0.2, 0.1, and 1.6 percent greater than observed load respectively). In addition, the correction of local deviations through Kalman filter or smoothing techniques in the F5, F7, and L7COMP methods (5.6, 5.6, and 13.5 percent greater than observed loads respectively for this site/sampling scenario) were more accurate than corresponding LOADEST estimates because these methods better utilize information gained from increased sample collection.

Load estimation method performance for TP and SSC differs substantially as compared to NO23 (Fig. 4). For TP and SSC, the RATIO method and the simpler forms of standard regression models (L5 and F5) have the lowest error levels on average with MPEs of 4% or less and RMSPEs of 21% or less. Regression-based methods utilizing quadratic flow and time terms (L7 and F7) tend

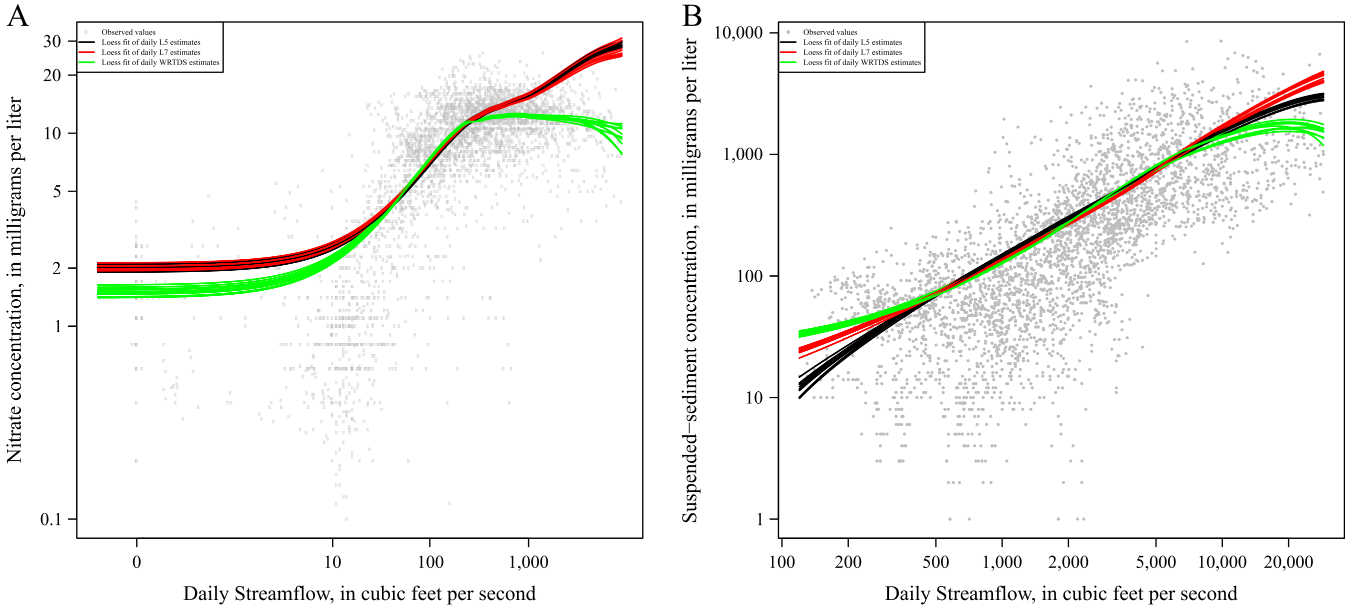


Fig. 5. (A and B) Examples of water-quality records in which LOADEST 5 and LOADEST 7 methods resulted in biased estimates. [(A) illustrates observed, daily values of nitrate at the Vermillion River at Pontiac, IL and loess fits of LOADEST 5 (L5), LOADEST 7 (L7) and WRTDS daily estimates for the non-furloughed, uniform sampling, 52 samples per year case. (B) illustrates observed, daily values of suspended sediment at the Skunk River at Augusta, IA and loess fits of LOADEST 5 (L5), LOADEST 7 (L7) and WRTDS daily estimates for the non-furloughed, uniform sampling, 52 samples per year case.]

to overestimate observed loads for TP and SSC on average. In many cases, regression-based methods severely over-estimate load due to misspecification of the concentration/discharge relation resulting in extremely large estimated concentration and load values during some high-discharge events. Those extreme over-estimates are reflected in the overall aggregate MPE and RMSPE values shown in Fig. 4. In Fig. 5B, the L7 model (and to a lesser degree the L5) overestimates SSC values at flows greater than about 5000 cubic feet per second; flows which account for 83 percent of decadal SSC load. The misspecification of concentration/discharge relations results in L7 estimates that are on average 46 percent biased at this site for this case (non-furloughed, uniform sampling, 52 samples per year) while L5 estimates are 25 percent biased. The accuracy of L5 estimates for SSC varied substantially among sampling sites, and were more frequently biased high for this particular site. As with NO23 estimates at the Vermillion River, RATIO and WRTDS estimates are less biased (−1.7 and −1.2 percent biased respectively) on average as compared to LOADEST estimates at this site. In comparison to L5 and L7 estimates, F5, F7, and L7COMP methods produce less-biased estimates on average (6, 9, and 4 percent respectfully) because of Kalman filter adjustments to local deviations from sampled values.

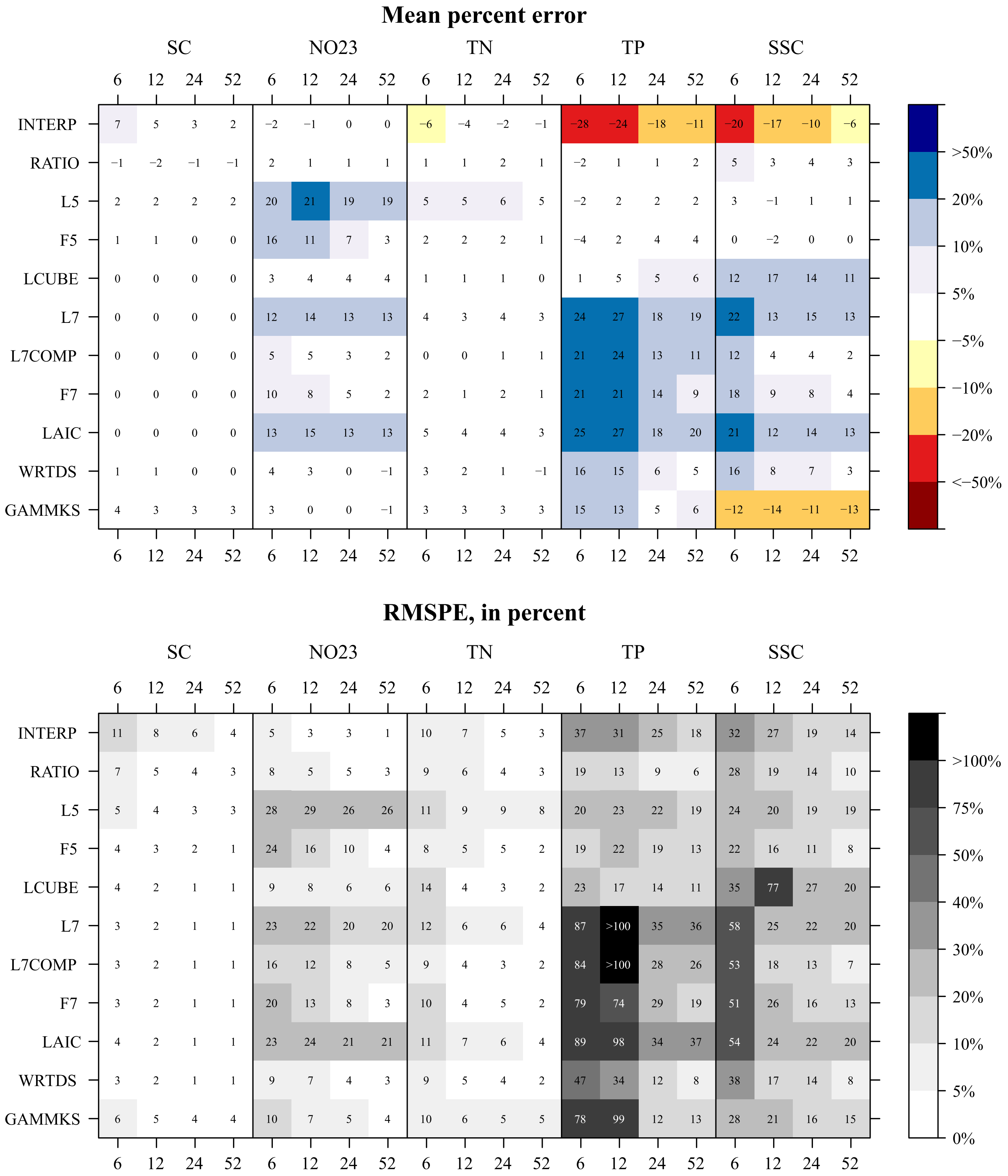
3.3. Effect of sampling frequency

Load estimation methods are evaluated over a range of sampling frequencies (6, 12, 24, and 52 samples per year) in order to detect improved performance with greater sampling rates. The results of these evaluations are presented by method and water-quality constituent aggregated over location and sampling strategy (Fig. 6 - low-flow and furloughed sample collection cases are omitted). Methods vary in the degree of improvement with increased sampling frequency. As observed among other scenarios, SC and TN estimates are on average within 5 percent of observed decadal loads and have smaller RMSPEs as compared to NO23, TP, and SSC estimates. For NO23, several of the regression methods (L5, F5, L7, F7, LAIC) have the highest aggregate MPE and RMSPE values. Local residual adjustments applied in some regression methods (L7COMP, F5, and F7) reduce MPE and RMSPE values compared to LOADEST (L5 and L7) estimates, especially as sampling frequency is increased. For TP and SSC, L5, F5, and RATIO provide estimates that have low bias (<5%) regardless of sampling frequency and WRTDS has low bias for the highest sampling frequency. In contrast, INTERP and methods that utilize quadratic representations of streamflow (L7, L7COMP, LAIC, F7) and GAMMKS provide TP and SSC estimates with the increased bias (greater than 10%). With regard to the RMSPE's for TP and SSC, the relative performance of the methods depend on sampling frequency. For lower frequencies (6 and 12 samples per year), L5, F5, and RATIO generally have the smallest RMSPEs and for higher frequencies (24 and 52 samples per year) RATIO, F5, WRTDS, and GAMMKS generally have the smallest RMSPE's. Interpolation (INTERP) provides TP and SSC load estimates that are consistently biased low because the method tends to underestimate transport during unsampled high-streamflow conditions that may persist for days to weeks, depending on sampling frequency.

More frequent sampling improves the performance of some estimation methods more than others. Correction for local departures from sampled values through Kalman filter or composite smoothing employed by INTERP, L7COMP, F5, and F7 methods substantially reduces RMSPEs for NO23, TP, and SSC relative to L5 and L7 estimates as sampling frequencies increase from 6 to 52 samples per year. Increased sampling frequency from 6 to 52 also reduces RMSPEs for WRTDS estimates by more than 60 percent for SC, NO23, and TN and more than 75 percent for TP and SSC. The smallest improvements with more frequent sampling (in terms of RMSPE) are generally observed for the L5 model. RMSPE values for L5 improve less than 30% from 6 to 52 samples per year for NO23, TP, and SSC.

3.4. Variation among sampling sites

Method performance is often affected by the environmental behavior of a specific constituent, as observed above and by the amount of hydrologic variability at a given monitoring site. These factors can lead to misspecification of estimation models and poor



Mean percent error

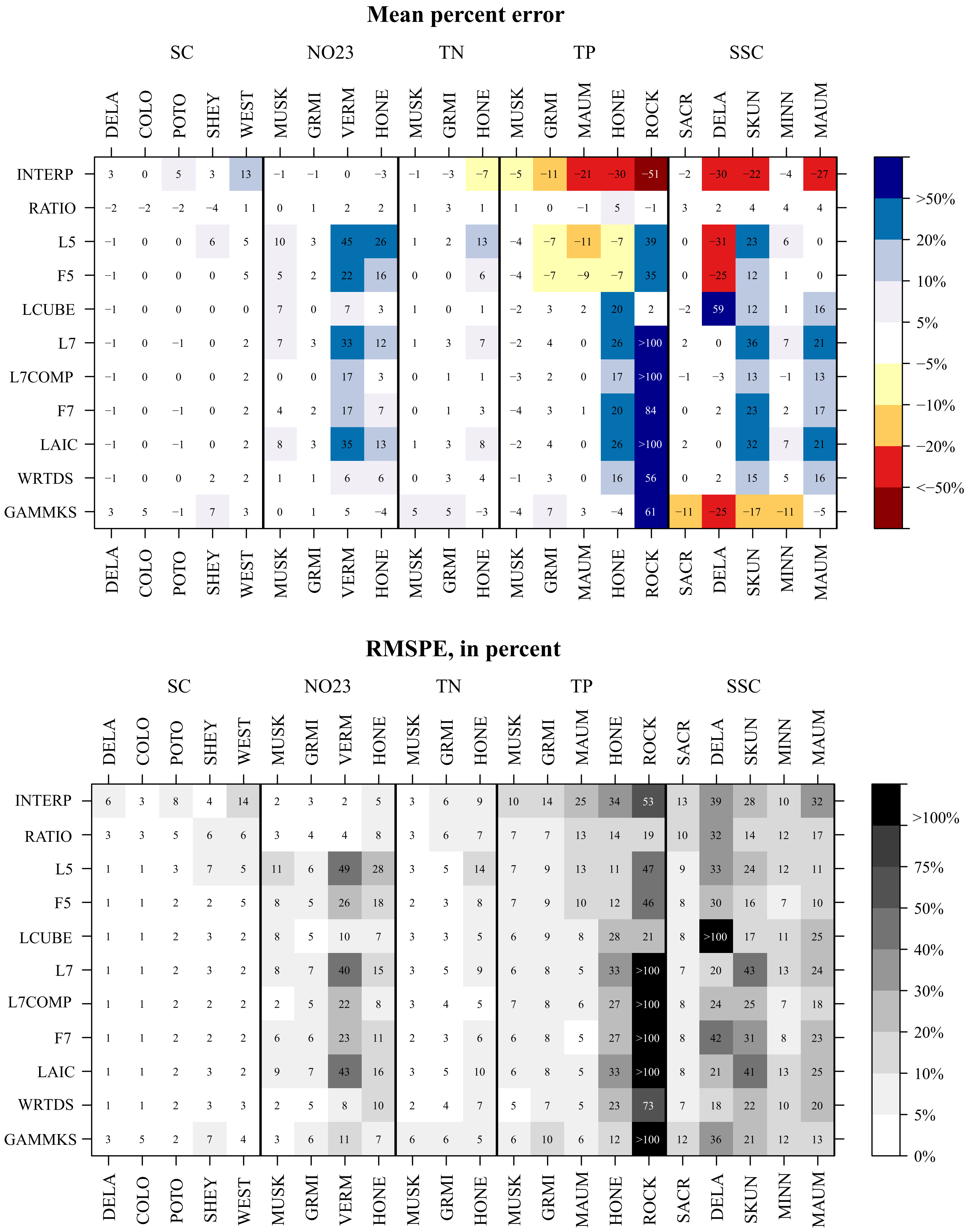
	SC				NO23				TN				TP				SSC			
	6	12	24	52	6	12	24	52	6	12	24	52	6	12	24	52	6	12	24	52
INTERP	7	5	3	2	−2	−1	0	0	−6	−4	−2	−1	−28	−24	−18	−11	−20	−17	−10	−6
RATIO	−1	−2	−1	−1	2	1	1	1	1	1	2	1	−2	1	1	2	5	3	4	3
L5	2	2	2	2	20	21	19	19	5	5	6	5	−2	2	2	2	3	−1	1	1
F5	1	1	0	0	16	11	7	3	2	2	2	1	−4	2	4	4	0	−2	0	0
LCUBE	0	0	0	0	3	4	4	4	1	1	1	0	1	5	5	6	12	17	14	11
L7	0	0	0	0	12	14	13	13	4	3	4	3	24	27	18	19	22	13	15	13
L7COMP	0	0	0	0	5	5	3	2	0	0	1	1	21	24	13	11	12	4	4	2
F7	0	0	0	0	10	8	5	2	2	1	2	1	21	21	14	9	18	9	8	4
LAIC	0	0	0	0	13	15	13	13	5	4	4	3	25	27	18	20	21	12	14	13
WRTDS	1	1	0	0	4	3	0	−1	3	2	1	−1	16	15	6	5	16	8	7	3
GAMMKS	4	3	3	3	3	0	0	−1	3	3	3	3	15	13	5	6	−12	−14	−11	−13

RMSPE, in percent

	SC				NO23				TN				TP				SSC			
	6	12	24	52	6	12	24	52	6	12	24	52	6	12	24	52	6	12	24	52
INTERP	11	8	6	4	5	3	3	1	10	7	5	3	37	31	25	18	32	27	19	14
RATIO	7	5	4	3	8	5	5	3	9	6	4	3	19	13	9	6	28	19	14	10
L5	5	4	3	3	28	29	26	26	11	9	9	8	20	23	22	19	24	20	19	19
F5	4	3	2	1	24	16	10	4	8	5	5	2	19	22	19	13	22	16	11	8
LCUBE	4	2	1	1	9	8	6	6	14	4	3	2	23	17	14	11	35	77	27	20
L7	3	2	1	1	23	22	20	20	12	6	6	4	87	>100	35	36	58	25	22	20
L7COMP	3	2	1	1	16	12	8	5	9	4	3	2	84	>100	28	26	53	18	13	7
F7	3	2	1	1	20	13	8	3	10	4	5	2	79	74	29	19	51	26	16	13
LAIC	4	2	1	1	23	24	21	21	11	7	6	4	89	98	34	37	54	24	22	20
WRTDS	3	2	1	1	9	7	4	3	9	5	4	2	47	34	12	8	38	17	14	8
GAMMKS	6	5	4	4	10	7	5	4	10	6	5	5	78	99	12	13	28	21	16	15

Fig. 6. Comparison of mean percent error and root mean squared percent errors for estimation methods among constituents and sampling frequencies relative to observed decadal loads. [SC, Specific conductance; NO23, Nitrate; TN, Total Nitrogen; TP, Total Phosphorus; SSC, Suspended sediment; Numbers indicate the number of samples per year; excludes load estimates from the low-flow and furlough sampling strategies.]

representation of concentration/discharge relations (Fig. 5). To assess the relative performance of estimation across a range of environmental conditions, we summarize error levels by constituent and by site (Fig. 7). MPE and RMSPE values are computed for each method, constituent, and site combination based on a "typical" sampling plan of one sample per month across the entire decade. These values are aggregated by replicate and across sampling strategies although the low-flow strategy and furloughs are



Mean percent error

	SC					NO23				TN			TP					SSC				
	DELA	COLO	POTO	SHEY	WEST	MUSK	GRMI	VERM	HONE	MUSK	GRMI	HONE	MUSK	GRMI	MAUM	HONE	ROCK	SACR	DELA	SKUN	MINN	MAUM
INTERP	3	0	5	3	13	−1	−1	0	−3	−1	−3	−7	−5	−11	−21	−30	−51	−2	−30	−22	−4	−27
RATIO	−2	−2	−2	−4	1	0	1	2	2	1	3	1	1	0	−1	5	−1	3	2	4	4	4
L5	−1	0	0	6	5	10	3	45	26	1	2	13	−4	−7	−11	−7	39	0	−31	23	6	0
F5	−1	0	0	0	5	5	2	22	16	0	0	6	−4	−7	−9	−7	35	0	−25	12	1	0
LCUBE	−1	0	0	0	0	7	0	7	3	1	0	1	−2	3	2	20	2	−2	59	12	1	16
L7	−1	0	−1	0	2	7	3	33	12	1	3	7	−2	4	0	26	>100	2	0	36	7	21
L7COMP	−1	0	0	0	2	0	0	17	3	0	1	1	−3	2	0	17	>100	−1	−3	13	−1	13
F7	−1	0	−1	0	2	4	2	17	7	0	1	3	−4	3	1	20	84	0	2	23	2	17
LAIC	−1	0	−1	0	2	8	3	35	13	1	3	8	−2	4	0	26	>100	2	0	32	7	21
WRTDS	−1	0	0	2	2	1	1	6	6	0	3	4	−1	3	0	16	56	0	2	15	5	16
GAMMKS	3	5	−1	7	3	0	1	5	−4	5	5	−3	−4	7	3	−4	61	−11	−25	−17	−11	−5

RMSPE, in percent

	SC					NO23				TN			TP					SSC				
	DELA	COLO	POTO	SHEY	WEST	MUSK	GRMI	VERM	HONE	MUSK	GRMI	HONE	MUSK	GRMI	MAUM	HONE	ROCK	SACR	DELA	SKUN	MINN	MAUM
INTERP	6	3	8	4	14	2	3	2	5	3	6	9	10	14	25	34	53	13	39	28	10	32
RATIO	3	3	5	6	6	3	4	4	8	3	6	7	7	7	13	14	19	10	32	14	12	17
L5	1	1	3	7	5	11	6	49	28	3	5	14	7	9	13	11	47	9	33	24	12	11
F5	1	1	2	2	5	8	5	26	18	2	3	8	7	9	10	12	46	8	30	16	7	10
LCUBE	1	1	2	3	2	8	5	10	7	3	3	5	6	9	8	28	21	8	>100	17	11	25
L7	1	1	2	3	2	8	7	40	15	3	5	9	6	8	5	33	>100	7	20	43	13	24
L7COMP	1	1	2	2	2	2	5	22	8	3	4	5	7	8	6	27	>100	8	24	25	7	18
F7	1	1	2	2	2	6	6	23	11	2	3	6	6	8	5	27	>100	8	42	31	8	23
LAIC	1	1	2	3	2	9	7	43	16	3	5	10	6	8	5	33	>100	8	21	41	13	25
WRTDS	1	1	2	3	3	2	5	8	10	2	4	7	5	7	5	23	73	7	18	22	10	20
GAMMKS	3	5	2	7	4	3	6	11	7	6	6	5	6	10	6	12	>100	12	36	21	12	13

Fig. 7. Comparison of mean percent error and root mean squared percent errors for estimation methods among constituents and sampling sites relative to observed decadal loads. [Column abbreviations indicated in Table 1; excludes load estimates from the low-flow and furlough sampling strategies.]

omitted, as described previously. Although environmental factors such as land-use and drainage area were examined, sites in Fig. 7 are ordered by constituent and then by increasing variability in streamflow (as defined by the coefficient of variation of daily streamflow - Table 1) from left to right.

Among sampling sites, load estimates tend to have higher error for sites with more variable streamflow conditions. In general, as in the previously described results, estimates for SC and TN have the lowest error with no method having an MPE greater than 13 percent in absolute value and most being less than 5 percent. However, even for these constituents, the sites with the greater flow variability, including the West Forks Stones River (WEST) for SC and Honey Creek (HONE) for TN, generally have the largest MPE and RMSPE values. For NO23, for all methods, the highest MPE and RMSPE values are observed at either the Vermillion River (VERM) or Honey Creek (HONE) sites, these sites having the most variable flows of the four considered. For TP, Rock Creek, which has by far the most variable flow conditions, has the highest MPEs for 9 of 11 methods, and has the highest RMSPEs for 10 of 11 methods. Unlike the other constituents, the error associated with SSC load estimates do not consistently increase with flow variability. The highest absolute MPE's and RMSPE's for most methods are observed at the Delaware River (DELA) and Skunk River (SKUN). The two sites with the highest streamflow variability for SSC (MINN and MAUM) happen to have large drainage areas as well (Table 1). Although there are too few sites to draw any definitive conclusions, it appears that drainage area, and probably other basin characteristics such as soil properties and channel morphology, play an important role in addition to streamflow variability for estimating SSC load.

Among the load estimation methods considered, some are more robust (less sensitive) to flow variability and other site-specific differences that may affect load estimates (Fig. 7). With respect to both MPE and RMSPE, the most robust method overall was RATIO, which has low bias (<5 percent) and low RMSPE (less than 20 percent) for all site/constituent combinations except one (SSC for DELA). None of the other methods have consistently low bias across all site/constituent combinations. However, for SSC, lower RMSPEs compared to RATIO could be obtained for 3 of 5 sites using alternative methods WRTDS, F5, and L7COMP. Furthermore, methods that involve fitting a model to daily loads, unlike RATIO, have the added advantage of being able to estimate (or simulate, depending on the application) loads or concentrations for specific days and flow conditions, evaluate temporal trends, estimate the probability of extreme events outside of the observed record, and other applications. The interpolation method (INTERP) consistently underestimates decadal load for constituents that exhibited positive concentration/discharge relations (TP and SSC) with the degree of underestimation often increasing with the amount of flow variability at sampling sites. Conversely, for NO23, INTERP is the best method in terms of both MPE and RMSPE, and the only method that is consistently better than RATIO. This is because nitrate concentrations at the four sites are generally less affected by changing streamflow conditions in comparison to other constituents. For example, in Fig. 5A, nitrate concentrations are not observed to consistently increase or decrease in response to changes discharge above about 100 cubic feet per second. In contrast, the RATIO method provides load estimates with low bias consistently across sites and seemed to be insensitive to the amount of flow variability at the sampling site. Among regression methods, those with quadratic or higher order representations of streamflow tend to overestimate load, particularly for sites with greater discharge variability. Regression methods with linear terms (L5 and F5) are among the few methods that underestimate TP and SSC load for multiple sites (GRMI, MAUM, and HONE for TP and DELA for SSC), but also overestimate load for selected sites (ROCK for TP and SKUN for SSC). However, it is important to recognize that the number of sites and the sets of conditions at those sites is limited in scope. Evaluation over a broader range of sites could help to better characterize method performance as a function of flow variability, high-flow event frequency and range, and other site characteristics that can potentially complicate load estimation.

3.5. Comparison of model performance relative to pre-defined criteria

While MPE and RMSPE values provide general comparisons of method performance, analysts often need to know if load estimates fall within pre-defined levels of accuracy. Constituent-specific, "acceptable" ranges of plus or minus 10 percent for TN and NO23 estimates, and plus or minus 20 percent for TP and SSC estimates are established for this purpose. Fig. 8 shows errors in decadal load estimates relative to these criteria for non-furloughed, uniform, seasonal, and high-flow estimates of NO23, TN, TP, and SSC; SC estimates are not shown because the magnitude of errors was very small among all of the methods. Grey areas indicate the acceptable range for each constituent. Table 2 indicates the percentage of estimates in Fig. 8 that fall within acceptable ranges of error, and ranks estimation method performance relative to criteria within and across selected constituents. This evaluation provides a different perspective on the relative performance of the load estimation methods. Aggregated MPE and RMSPE values summarized in the level plots could be dominated by a limited number of cases with extreme errors. While that information is still useful, it does not describe the frequency with which methods provide an acceptable result. Thus Fig. 8 and Table 2 are included to provide some information in that regard. Additional evaluations of the robustness of methods relative to specified criteria among sampling furloughs and frequencies is included in the supplemental information.

Methods with the ability to flexibly define relations among constituent concentrations and flow conditions (RATIO, WRTDS, and GAMMKS) produce the most estimates within specified criteria. RATIO produced the most estimates within criteria among all constituents and produced no estimates that are extremely biased estimates such as those observed for methods applied for selected TP and SSC cases (Fig. 8). RATIO TN estimates are within criteria less often than many regression-based methods because TN loads are generally well represented by linear, quadratic, or cubic representations of flow conditions. WRTDS estimates are within the top four among methods for each constituent with respect to criteria, and similar to RATIO results, the median estimates are nearly unbiased for each constituent tested. GAMMKS estimates are within the top 4 methods for TN, NO23, and TP, but tend to underestimate SSC loads.

Methods that provide the ability to adjust for local departures from actual measurements, such as L7COMP, F5 and F7, provide more estimates within criteria relative to L5 and L7 models; and have the 4th, 7th, and 6th most estimates within criteria respectively. Among LOADEST methods, cubic representations of flow provide more estimates within criteria among all constituents, while quadratic representations of flow in the L7 model provides more estimates within criteria relative to L5 for NO23, TN, and SSC. It should be noted however that many of the regression methods with higher order terms (cubic and quadratic) produced some estimates with severe error, a result also reported by Hirsch (2014). The use of Akaike's Information Criterion to select independent variables does not improve performance over L7 or LCUBE methods. Contrary to more favorable aggregate results observed in Figs. 2–4, 6 and 7, the L5 model results in the fewest estimates within criteria across all constituents. It is important to note that the comparison of estimates to criteria offer only one perspective on method performance, and that although L5 results fall outside criteria the most frequently, this method avoids the extreme bias

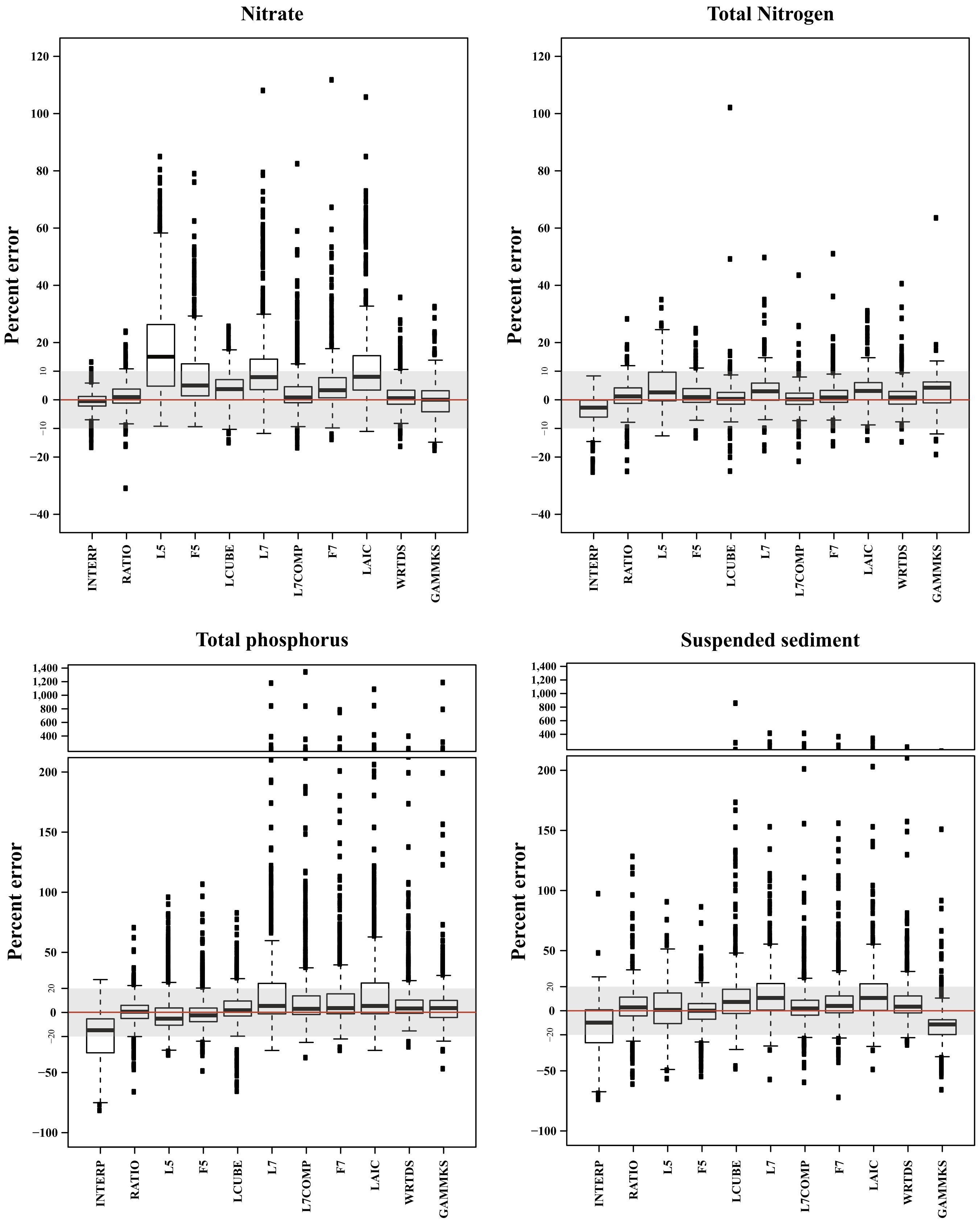


Fig. 8. Percentage errors and for decadal nitrate, total nitrogen, total phosphorus and suspended-sediment load for estimation methods relative to observed decadal loads. [Estimates obtained from samples collected under non-furloughed, uniform, seasonal, and high-flow sampling scenarios.]

Table 2
Percent of load estimates within specified tolerances of true loads. [Results exclude estimates from low-flow or furlough sampling strategies.]

	NO23 (±10%)	NO23 rank	TN (±10%)	TN rank	TP (±20%)	TP rank	SSC (±20%)	SSC rank	Rank among all constituents
INTERP	97.5	1	89.4	8	58.2	11	65.0	11	8
RATIO	92.9	2	91.1	7	90.2	1	86.2	1	1
L5	38.1	11	75.3	11	81.0	6	68.3	10	11
F5	68.8	8	92.2	5	83.3	5	85.0	2	7
LCUBE	84.0	5	94.4	3	83.8	4	75.5	6	4
L7	63.3	9	87.5	9	70.8	9	68.5	9	9
L7COMP	82.1	6	95.0	1	78.7	7	82.0	3	4
F7	81.9	7	92.2	5	77.7	8	79.7	5	6
LAIC	61.7	10	86.9	10	70.8	9	68.7	8	10
WRTDS	89.4	3	94.2	4	84.7	3	81.3	4	2
GAMMKS	89.0	4	95.0	1	87.8	2	72.3	7	3

for selected cases observed with some other methods (Fig. 8). Although INTERP is among the worst performing methods for TN, TP, and SSC estimates, it is the best performing method for NO23, with only 2 percent of estimates falling outside of 10 percent of observed loads. These results indicate that at the sites tested, NO23 concentrations do not vary substantially within a particular month or relative to flow conditions, thus allowing a simple method like INTERP to provide decadal load estimates that are consistently within the set criteria.

4. Discussion

Load estimation method performance is dependent upon a variety of factors including constituent type, streamflow characteristics, sampling strategy, sampling frequency, and water-quality record consistency. Ideally a single load estimation method could be identified to consistently provide accurate estimates across all of these factors. However, none of the methods we consider are fully robust across the full range of record characteristics considered in this study, although some are more consistently accurate than others.

One factor that clearly affects the level of error in load estimation methods is water-quality constituent type and the environmental behavior exhibited by a specific constituent at a specific site. Of the constituents that we use as test cases, those that exhibit relatively consistent concentration/discharge relations (specific conductance and total nitrogen) have load estimates with low error no matter which method is used (usually less than 10 percent). In these cases, most methods work well and a method can be selected based on other factors such as the need to simultaneously remove trends. In contrast, those constituents that exhibit strongly positive concentration/discharge relation (total phosphorus and suspended sediment) are sometimes poorly estimated. Substantial curvature in the log concentration to log discharge relation can lead to estimates that are severely biased. Methods that use quadratic or cubic representations of log streamflow are particularly problematic especially when sample data sets do not contain a substantial number of high discharge samples. These results indicate that the constituent type should be a primary consideration when selecting a method for load estimation and particular care should be used in both selecting and implementing a method for estimating loads of those constituents that are transported mainly during high-discharge events. If regression methods are used for such constituents, they should be applied with caution, making extensive use of diagnostic methods as recommended by Hirsch (2014).

In contrast to the other four water-quality constituents, nitrate presents some unique challenges that cause many regression methods to over-estimate decadal load. Nitrate is often highly variable during the warmer months of the year due to seasonal applications of fertilizer, variations in patterns of transport through shallow ground water, and the efficacy of denitrification. Nitrate can increase rapidly either during hydrograph rises or after high-discharge events. These patterns may cause heteroscedasticity and are discussed by Stenback et al. (2011) and Hirsch (2014) who noted their effects on log-retransformation bias correction typically employed using regression techniques. In addition, relations of constituent concentrations with flow often are not accurately specified through linear, quadratic, or cubic representations of flow conditions. Simple methods like interpolation and the ratio estimator that do not rely on bias correction methods or attempt to statistically model the relation between flow and concentration often provide nitrate load estimates that are more accurate. Consistent with the results of Hirsch (2014) more flexible regression methods like WRTDS and GAMMKS are generally better able to estimate decadal nitrate loads than LOADEST or FLUXMASTER regression methods. Thus these latter techniques may provide a better option for estimating nitrate loads and caution should be utilized when employing traditional regression techniques for that purpose.

No matter which method is selected for load estimation, water-quality data records developed using well designed and consistent sampling programs will help to minimize error. Based on the results summarized in Fig. 3, the importance of sampling during high-discharge periods is evident. Neglecting high-discharge entirely will elevate error levels in load estimates no matter what method is used and no matter what constituent is considered. Some high-discharge sampling through uniform or seasonal sampling programs will improve load estimate accuracy, but targeted high-discharge sampling offers the greatest potential for minimizing error. In general increased sampling frequency offered the most improvement when using interpolation, newer methods (WRTDS and GAMMKS), and methods that use residual smoothing techniques (L7COMP, F5, and F7) to estimate nitrate, total phosphorus, or suspended-sediment loads.

Variation in funding levels can cause monitoring program cutbacks that result in gaps in the monitoring records where long-term data collection has existed. Based on our results (summarized in Fig. 4), these gaps, which we call “furloughs”, can have a substantial impact on the accuracy of some load estimation methods. Many of the traditional regression methods with quadratic representations of discharge are most susceptible to error introduced by furloughs. This was particularly true for those constituents driven by high-discharge events (total phosphorus and suspended sediment); where these methods have decreased accuracy when furloughs are present in the record. Traditional regression techniques are often applied in cases where furloughs are present with the assumption that they could be used to interpolate across time periods when no data were collected. However, our results indicate that load estimate error can be inflated substantially when traditional regression methods are applied using records with temporal gaps. Thus caution is recommended when load estimation is required for records with periods of missing water-quality measurements, especially when missing records necessitate extended extrapolation through time.

Comparisons among estimation methods indicate that the flexibility in defining relations between load and flow conditions inherent in RATIO, WRTDS, and GAMMKS methods can improve the accuracy of decadal load estimates relative to more strictly-defined regression methods, particularly for nitrate, total phosphorus and suspended-sediment. Among more traditional regression approaches, residual smoothing techniques offered by FLUXMASTER-K and the composite method (L7COMP) generally improve load estimates relative to unadjusted LOADEST methods. Among LOADEST models considered, cubic representation of flow conditions can increase the likelihood of providing accurate estimates, but can also lead to extreme errors for selected cases due to misspecification of relations among flow and constituent concentration. The use of Akaike's Information Criterion to select independent variables did not consistently improve the accuracy of load estimates relative to LOADEST 5 and LOADEST 7 models. Although interpolation provides relatively unbiased estimates of nitrate load for the sites considered, it is generally inadequate for other constituents, especially in cases with relatively infrequent sampling.

Of the methods considered, the ratio estimator (RATIO) was the most consistent in providing accurate and unbiased decadal load estimates. Only in the case of low-flow sampling does the ratio estimator provide load estimates with average bias greater than 10 percent. These results are consistent with previous studies (Dolan et al., 1981; Richards and Holloway, 1987) and confirm the value of the ratio estimator for load estimation. A limitation of the ratio estimator is that it does not provide an ability to perform other types of analyses that are associated with and often reliant on the method used for load estimation. For example, analyses of trends in load require the definition of relations of load with both time and discharge, and many of the regression techniques are designed with that purpose in mind. Similarly SPARROW models require load estimates that are normalized to a given point in time to facilitate comparison of loads from monitoring records covering different periods, which requires knowledge of the effects of trend on loads, both through the direct effect holding flow constant and through the indirect effect of trending flow; the FLUXMASTER-K and WRTDS models were designed to provide that ability. Despite these limitations, knowledge of the robustness of the ratio estimator with regard to bias is valuable for load estimation and could be valuable as a diagnostic measure for the evaluation of load estimates using other methods.

Before choosing among estimation methods, it is important that analysts first identify characteristics specific to their particular application, such as the constituents being estimated, the size and flow variability of the sampling sites, whether estimates need to be computed via an automated method (i.e. without inspection of residual plots), the amount of error permissible, or the need to normalize results with respect to time or flow conditions. In addition, it is important to understand that while some methods may compute accurate estimates on average (such as L5), the individual estimates may be less likely to fall within a specified level of error. When possible, analysts should inspect diagnostics and plots illustrating method fit to observed data prior to estimating load (see Hirsch, 2014). In cases where estimates must be computed via an automated method and there is a need to de-trend or flow-normalize estimates, quantifying the departure of load estimates from those computed by the RATIO method can provide a useful indication of potential bias.

The ability of existing methods to estimate decadal loads of selected water-quality constituents is one aspect of a larger problem. Additional constituents, such as metals, pesticides, and other organic constituents can be transported in ways that existing methods are not adequate to represent. Many applications require load estimation for an individual year, season, or month; these are periods in which estimates are sensitive to specific hydrologic or seasonal conditions. In addition, the accuracy of existing methods to estimate the precision of load estimates is not known. Future efforts are needed to assess the ability of existing methods to estimate the standard error of load estimates, and to assess approaches to compute constituent load over shorter time spans.

5. Conclusions

Tests of a variety of new and commonly-used estimation methods indicate that no single method always produces the most accurate decadal load estimates among different constituents, sites, and sampling scenarios. Sampling record characteristics can affect the accuracy of load estimates and these results emphasize the need for case-by-case evaluations of method fit to avoid load estimate bias. However, for applications requiring automated estimation, most methods work well for constituents related to specific conductance and for total nitrogen, whereas methods that allow for flexibility in relations between streamflow and load, such as Weighted Regression on Time, Season, and Discharge (WRTDS) and the Beale Ratio estimator, are most likely to provide relatively accurate estimates of nitrate, total phosphorus, and suspended-sediment. Kalman smoothing methods demonstrate utility in improving the accuracy of load estimation, at least for higher sampling frequencies. In cases where there is a need to de-trend or flow-normalize load estimates, comparison of the un-normalized estimates to the Beale's Ratio estimate may provide a useful indicator of bias. Additional work is needed to evaluate existing methods for estimating the error of load estimates, to identify metrics that might indicate the likelihood that a method will provide an accurate estimate, and to characterize the bias and variability of existing methods to estimate the load of water-quality constituents at shorter time spans such as years, seasons, and months. Future work could also include evaluating the potential of in-situ water-quality sensors to improve the accuracy of load estimates at multiple time scales.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jhydrol.2016.08.059.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automatic Control. 19 (6), 716–723.
Aulenbach, B.T., Hooper, R.P., 2006. The composite method: an improved method for stream-water solute load estimation. Hydrol. Process. 20, 3029–3047.
Cochran, W.G., 1977. Sampling Techniques. Wiley, New York.
Cohn, T.A., DeLong, L.L., Gilroy, E.J., Hirsch, R.M., Wells, D.K., 1989. Estimating constituent loads. Water Resour. Res. 25 (5), 937–942.
Cohn, T.A., Caulder, D.L., Gilroy, E.J., Zynjuk, L.D., Summers, R.M., 1992. The validity of a simple statistical model for estimating fluvial constituent loads: an empirical study involving nutrient loads entering Chesapeake Bay. Water Resour. Res. 28 (9), 2353–2363.
Cohn, T.A., 2005. Estimating contaminant loads in rivers: an application of adjusted maximum likelihood to type 1 censored data. Water Resour. Res. 41 (7), 1–13.
Crawford, C.G., 1991. Estimation of suspended-sediment rating curves and mean suspended-sediment loads. J. Hydrol. 129, 331–348.
Dolan, D.M., Yui, A.K., Geist, R.D., 1981. Evaluation of river load estimation for total phosphorus. J. Great Lakes Res. 7 (3), 207–214.
Falcone, J.A., 2011. GAGES-II, Geospatial attributes of gages for evaluating streamflow [digital spatial dataset]. At <http://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml> (accessed on July 14, 2016).

Ferguson, R.I., 1986. River loads underestimated by rating curves. Water Resour. Res. 22 (1), 74–76.

Ferguson, R.I., 1987. Accuracy and precision of methods for estimating river loads. Earth Surf. Process. Landf. 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.

Heidelberg University, 2005. User's Guide to the River Data Sets, Chapter 1b, Sampling Stations and Methods. At <http://www.heidelberg.edu/academiclife/distinctive/ncwqr/data/guide> (accessed on July 14, 2016).

Hem, J.D., 1985. Study and interpretation of the chemical characteristics of natural water. U.S. Geological Survey Water Supply Paper 2254.

Hirsch, R.M., Moyer, D.L., Archfield, S.A., 2010. Weighted Regressions on Time, Discharge, and Season (WRTDS), with an application to Chesapeake Bay River inputs. J. Am. Water Resour. Assoc. 46 (5), 857–880.

Hirsch, R.M., 2014. Large biases in regression-based constituent load estimates: causes and diagnostic tools. J. Am. Water Resour. Assoc. 50 (6), 1401–1424.

Hirsch, R.M., De Cicco, L.A., 2014. User Guide to Exploration and Graphic for RivEr Trends (EGRET) and dataRetrieval: R Packages for Hydrologic Data. U.S. Geological Survey Techniques and Methods 4-A10.

Horowitz, 2003. An evaluation of sediment rating curves for estimating suspended sediment concentration for subsequent load calculations. Hydrol. Process. 17, 3387–3409.

Lee, C.J., Glysson, G.D., 2013. Compilation, quality control, analysis, and summary of discrete suspended-sediment and ancillary data in the United States, 1901–2010. U.S. Geological Survey Data Series 776.

Moyer, D.L., Hirsch, R.M., Hyer, K.E., 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.

O'Connor, D., 1976. The concentration of dissolved solids and river flow. Water Resour. Res. 12 (12), 279–294.

Preston, S.D., Bierman Jr., V.J., Sillman, S.E., 1989. An evaluation of methods for the estimation of tributary mass loads. Water Resour. Res. 25, 1379–1389.

Richards, R.P., Holloway, J., 1987. Monte Carlo studies of sampling strategies for estimating tributary loads. Water Resour. Res. 23 (10), 1939–1948.

Richards, R.P., Alameddine, I., Allan, J.D., Baker, D.B., Bosch, N.S., Confesor, R., DePinto, J.V., Dolan, D.M., Reutter, J.M., Scavia, D., 2012. Discussion: nutrient inputs to the Laurentian Great Lakes by source and watershed estimated using SPARROW watershed models. J. Am. Water Resour. Assoc. 49 (3), 715–724.

Robertson, D.M., Roerish, E.D., 1999. Influence of various water quality sampling strategies on load estimates for small streams. Water Resour. Res. 35 (12), 3747–3759.

Robertson, D.M., 2003. Influence of different temporal sampling strategies on estimating total phosphorus and suspended sediment concentration and transport in small streams. J. Am. Water Resour. Assoc. 39 (25), 1281–1310.

Runkel, R.L., Crawford, C.G., Cohn, T.A., 2004. Load Estimator (LOADEST): A FORTRAN Program for Estimating Constituent Loads in Streams and Rivers. U. S. Geological Survey Techniques and Methods Book 4, Chapter A5.

Ryberg, K.R., Vecchia, A.V., 2012. waterData—An R Package for Retrieval, Analysis, and Anomaly Calculation of Daily Hydrologic Time Series Data, Version 1.0. U.S. Geological Survey Open-File Report 2012-1168.

Saad, D.A., Schwarz, G.E., Robertson, D.M., Booth, N.L., 2011. A multi-agency nutrient dataset used to estimate loads, improve monitoring design, and calibrate regional nutrient SPARROW models. J. Am. Water Resour. Assoc. 47 (5), 933–949.

Schwarz, G.E., Hoos, A.B., Alexander, R.B., Smith, R.A., 2006. The SPARROW Surface Water-Quality Model—Theory, Applications and User Documentation. U.S. Geological Survey Techniques and Methods, Book 6, Chapter B3.

Stenback, G.A., Crumpton, W.A., Schilling, K.E., Helmers, M.J., 2011. Rating curve estimation of nutrient loads in Iowa Rivers. J. Hydrol. 396, 158–169.

Tesoriero, A.J., Duff, J.H., Saad, D.A., Spahr, N.E., Wolock, D.M., 2013. Vulnerability of streams to legacy nitrate sources. Environ. Sci. Technol. 38, 1892–1900.

Verma, S., Markus, M., Cooke, R.A., 2012. Development of error correction techniques for nitrate-N load estimation methods. J. Hydrol. 432–433, 12–25.

Walling, D.E., 1977. Assessing the accuracy of suspended sediment rating curves for a small basin. Water Resour. Res. 13 (3), 531–538.

Wolman, M.G., Miller, J.P., 1960. Magnitude and frequency of forces in geomorphic processes. J. Geol. 68, 54–74.