USGS
science for a changing world

An Evaluation of Methods for Computing Annual Water-Quality Loads

Scientific Investigations Report 2019–5084

U.S. Department of the Interior
U.S. Geological Survey

An Evaluation of Methods for Computing Annual Water-Quality Loads

By Casey J. Lee, Robert M. Hirsch, and Charles G. Crawford

Scientific Investigations Report 2019–5084

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
DAVID BERNHARDT, Secretary

U.S. Geological Survey
James F. Reilly II, Director

U.S. Geological Survey, Reston, Virginia: 2019

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment—visit https://www.usgs.gov or call 1–888–ASK–USGS.

For an overview of USGS information products, including maps, imagery, and publications, visit https://store.usgs.gov.

Suggested citation:
Lee, C.J., Hirsch, R.M., and Crawford, C.G., 2019, An evaluation of methods for computing annual water-quality loads: U.S. Geological Survey Scientific Investigations Report 2019–5084, 59 p., https://doi.org/10.3133/sir20195084.

ISSN 2328-0328 (online)

Foreword

Sustaining the quality of the Nation's water resources and the health of our diverse ecosystems depends on the availability of sound water-resources data and information to develop effective, science-based policies. Effective management of water resources also brings more certainty and efficiency to important economic sectors. Taken together, these actions lead to immediate and long-term economic, social, and environmental benefits that make a difference in the lives of the almost 400 million people projected to live in the United States by 2050.

In 1991, Congress established the National Water-Quality Assessment (NAWQA) Program (https://water.usgs.gov/nawqa/applications/) to address where, when, why, and how the Nation's water quality has changed, or is likely to change in the future, in response to human activities and natural factors. Since then, NAWQA has been a leading source of scientific data and knowledge used by national, regional, State, and local agencies to develop science-based policies and management strategies to improve and protect water resources used for drinking water, recreation, irrigation, energy development, and ecosystem needs. Plans for the third cycle of NAWQA (2013–21) address priority water-quality issues and science needs identified by NAWQA stakeholders, such as the Advisory Committee on Water Information and the National Research Council, and are designed to meet increasing challenges related to population growth, increasing needs for clean water, and changing land-use and weather patterns.

Federal, State, and local agencies have invested billions of dollars to reduce the amount of pollution entering rivers and streams that millions of Americans rely on for drinking water, recreation, and irrigation. Accurate information on the loading of water-quality constituents is crucial for evaluating the effectiveness of pollution control efforts and protecting the Nation's water resources into the future. This report helps to improve these methods through an evaluation of methods for computing annual water-quality loads at water-quality sampling sites. All NAWQA reports are available online (https://water.usgs.gov/nawqa/bib/).

We hope this publication will provide you with insights and information to meet your water-resource needs and will foster increased citizen awareness and involvement in the protection and restoration of our Nation's waters. The information in this report is intended primarily for those interested or involved in resource management and protection, conservation, regulation, and policymaking at the regional and national levels.

Dr. Donald W. Cline
Associate Director for Water
U.S. Geological Survey

Contents

Figures

Tables

Conversion Factors

U.S. customary units to International System of Units

Multiply	By	To obtain
Flow rate		
cubic foot per second (ft^3/s)	0.02832	cubic meter per second (m^3/s)

Supplemental Information

A water year is the period from October 1 to September 30 and is designated by the year in which it ends; for example, water year 2015 was from October 1, 2014, to September 30, 2015.

Concentrations of chemical constituents in water are given in milligrams per liter (mg/L).

Abbreviations

AIC	LOADEST minimum Akaike information criteria method with additional explanatory variables
AIC_COMP	LOADEST minimum Akaike information criteria method with additional explanatory variables and adjustment via the composite method
BIMONTH	bimonthly sampling strategy
BIWEEK	biweekly sampling strategy
HIFLOW	high-flow sampling strategy
HIFLOWE	high-flow early sampling strategy
INTERP	interpolation of sampled values method
L1	LOADEST stock 1-parameter model with streamflow as the only explanatory variable
L5	LOADEST stock 5-parameter model with streamflow, season, and time as explanatory variables
L7	LOADEST stock 7-parameter model with streamflow, streamflow squared, season, time, and time squared as explanatory variables
LAICO	LOADEST stock “best selection” model that selects explanatory variables with the minimum Akaike information criteria
MONTH	monthly sampling strategy
MSE	mean-squared error
NAWQA	U.S. Geological Survey National Water-Quality Assessment Program
NWQN	U.S. Geological Survey National Water Quality Network
PVAL	LOADEST minimum probability values method with additional explanatory variables

RATIO_F1	Beale's ratio estimation with streamflow-based stratification
RATIO_F5	Beale's ratio estimation with streamflow-based stratification on the most recent 5 years of data
RATIO_T	Beale's ratio estimation with time-based stratification
USGS	U.S. Geological Survey
WRTDS	Weighted Regressions on Time, Discharge, and Season method
WRTDS_K	Weighted Regressions on Time, Discharge, and Season method with Kalman filtering

An Evaluation of Methods for Computing Annual Water-Quality Loads

By Casey J. Lee, Robert M. Hirsch, and Charles G. Crawford

Abstract

The U.S. Geological Survey publishes information on the mass, or load, of water-quality constituents transported through rivers and streams sampled as part of the operation of the National Water Quality Network (NWQN). This study evaluates methods for computing annual water-quality loads, specifically with respect to procedures currently (2019) used at sites in the NWQN. Near-daily datasets of chloride, total nitrogen, nitrate plus nitrite, total phosphorus, and suspended sediment were subset to determine the accuracy of various load-estimation methods, including linear interpolation, ratio estimators, and linear and weighted-regression methods. Water-quality loads are computed under different sampling strategies and at multiple sampling sites to provide a more complete evaluation of load-estimation methods.

Estimation methods were less accurate when computing loads at annual rather than decadal time steps. Depending on the water-quality constituent, annual loads were within comparable accuracy thresholds 21 to 64 percent of the time relative to decadal loads. The accuracy of annual load estimates varied among water-quality constituents, sampling strategies, sampling sites, and estimation methods. Methods were most accurate when estimating chloride and decreased in accuracy when estimating total nitrogen, nitrate plus nitrite, total phosphorus, and suspended-sediment loads. Estimation methods were most likely to compute accurate annual loads when samples were collected frequently (26 samples per year) and when sampling strategies targeted high-flow conditions. For a given water-quality constituent, estimation accuracy differed substantially among sampling sites; estimates were more likely to be accurate at large rivers with less variability in concentration and (or) discharge conditions and were less likely to be accurate at smaller stream sites with more variable streamflow and (or) water-quality concentrations.

The Weighted Regressions on Time, Discharge, and Season method with Kalman filtering (WRTDS_K) generally produced the most accurate annual load estimates among sampling sites and water-quality constituents. Although WRTDS_K was the most accurate generally, every estimation method evaluated had the potential to produce accurate (and inaccurate) load estimates depending on the site, constituent, and water year. Linear interpolation and ratio estimators that used samples exclusively from the year being estimated were among the best performing methods for total nitrogen and nitrate plus nitrite loads but were among the least accurate when estimating annual total phosphorus and suspended-sediment loads. Ratio estimation that considered samples from previous years and stratified based on streamflow conditions produced among the most accurate total phosphorus estimates but was among the least accurate for other constituents. Regression-based methods that assumed linear or quadratic relations among the logarithm of water-quality concentrations and streamflow conditions were among the least accurate methods generally, whereas regression-based methods that considered cubic relations among the logarithm of concentration and streamflow and the Weighted Regressions on Time, Discharge, and Season (WRTDS) method were typically more accurate. Methods that adjusted daily estimates computed from regression or weighted-regression methods based on departures from sampled values, such as WRTDS_K and the composite method, improved estimate accuracy for most sites and constituents, but especially for chloride, total nitrogen, nitrate plus nitrite, and suspended-sediment estimates.

Investigation of the underlying causes of estimation method bias indicated that sites and years with more variability in concentration and loading conditions, higher slopes in the relation of the logarithm of concentration and discharge, and sampling plans that underrepresented high-flow conditions generally led to less accurate load estimates. Finally, because all methods indicated the capacity to produce biased load estimates, additional work is needed to identify the capacity of new technologies, such as continuous water-quality sensors, to improve the accuracy of annual or shorter term load estimates. Based on findings in this report, the NWQN will continue to publish water-quality loads using LOADEST-based methods that consider multiple transformations of streamflow, as well as season, time, and variables indicative of historical streamflow conditions to maintain consistent methods for stakeholders. However, the NWQN also plans to begin publishing annual load estimates using the WRTDS_K method in 2020 because this method was determined to be the most accurate for a given site, constituent, and water year.

Introduction

Knowledge of the mass, or load, of water-quality constituents transported by streams and rivers is necessary to assess the health of receiving waters and to characterize contributions from upstream landscapes. Load is expressed as the total mass of a water-quality constituent passing a stream location over a given time step, such as a day, year, or decade. Water-quality loads are quantified by summing the product of streamflow and water-quality constituent concentrations at frequent (that is, 15-minute to daily) time steps. Streamflow is quantified through frequent, automated collection of stage measurements; periodic stream discharge measurements; and the calibration of stage/discharge relations. The expense and time required to obtain instream samples generally dictates that water-quality data are available at monthly or less frequent time steps. Thus, to quantify water-quality loads, methods must be used to estimate water-quality concentrations on days when no samples are collected.

Several methods have been used to estimate water-quality loads within and outside of the U.S. Geological Survey (USGS). These methods include simple interpolation techniques, ratio estimators (Cochran, 1977), regression-based techniques (Ferguson, 1986; Cohn and others, 1989; Cohn and others, 1992), and more recently, a weighted-regression technique, Weighted Regressions on Time, Discharge, and Season (WRTDS), which is designed to account for the changing nature of relations between streamflow and water-quality constituents with respect to time and season (Hirsch and others, 2010; Hirsch and others, 2015).

Recent studies by Stenback and others (2011) and Richards and others (2012) highlighted the potential for regression-based methods, such as those used within the USGS LOADEST program (Runkel and others, 2004), to produce highly biased estimates when applied without careful scrutiny. Hirsch (2014) evaluated WRTDS and two LOADEST model configurations (with and without streamflow and time-squared terms) for nitrate and total phosphorus loads at three Midwest sites and generally determined that WRTDS offered more accurate estimates, although there were still cases in which it produced biased results. Lee and others (2016) evaluated the accuracy of 11 methods for computing decadal water-quality loads and generally determined that methods that allowed for flexibility in determining concentration and discharge relations, such as ratio estimators or WRTDS, produced the most accurate loads. However, the accuracy of decadal load estimates in this study varied substantially across constituents, sites, and sampling conditions.

National-scale USGS networks have used a variety of methods to compute loads from data collected at long-term monitoring stations (Lee and others, 2017a). Currently (2019), loads are computed at USGS National Water Quality Network (NWQN) sites using an adapted-LOADEST method that uses water-quality and streamflow data obtained from a 5-year moving window (Lee and others, 2017a). The adapted-LOADEST method uses additional explanatory variables not included in the default model choices provided in the original LOADEST program (Runkel and others, 2004) and includes an additional step that forces an analyst to inspect the fit of candidate models through a series of graphs before publication (Deacon and others, 2015; Lee and others, 2017a). Previous evaluations of load-estimation procedures have had little application to USGS NWQN operations because (1) they usually are not focused on annual time steps and (2) they typically do not evaluate sampling strategies and estimation methods used by the USGS NWQN.

Purpose and Scope

The purpose of this publication is to expand upon results presented in Lee and others (2016) to evaluate methods for computing water-quality loads at an annual time step, with additional consideration of methods used at USGS NWQN sites. This report considers previously untested estimation methods and examines the underlying causes of estimation method bias. Results can help practitioners inside and outside the USGS understand when, and to what degree, various sampling procedures and load-estimation methods are likely to produce accurate water-quality load estimates at an annual time step.

Methods

Estimation methods are evaluated by (1) obtaining data from sites with long-term, daily records of constituent concentrations; (2) subsetting acquired daily records using various sampling strategies; (3) estimating annual loads from these subsets using different methods; and (4) comparing estimated annual loads to the sum of observed data for a given site, water-quality constituent, and water year. The estimation methods considered range from simple to relatively complex and include simple interpolation, various iterations of ratio estimators, various forms of simple and multiple regression (implemented through the USGS LOADEST program), and weighted regression (implemented through WRTDS).

Load-estimation methods in this study have different strategies for considering data from years prior to the year being estimated (hereafter referred to as the “target year”). Some methods are designed to use data from the target year exclusively, some may use data from all years up to and including the target year, and others may use data from a specified number of years up to and including the target year. In this study, we generally use a fixed-window length of 5 years for methods in this latter category because this is the approach used to estimate loads as part of the USGS NWQN (Lee and others, 2017a). The use of a 5-year window means that on the fifth year of a given water-quality sampling record, data from the fifth year and the previous 4 years are used to compute the annual load for the target year. The assumption behind

this approach is that practitioners are computing loads in real time, do not have access to future water-quality observations, and do not alter previously computed loads as new data are collected because stakeholders often prefer results that do not change from year to year. However, it is important to note that most methods considered in this report could use any number of water-quality samples and could be applied in a manner in which target-year estimates are revised as additional data are collected beyond the target year. Because the use of a 5-year window is largely arbitrary, a specific analysis is described in appendix 5 to evaluate the accuracy of different "sampling window" lengths. See the "Load Estimation Methods" section for more information on how various methods use historical water-quality samples.

Datasets Used for this Evaluation

Daily observations of water-quality concentration and streamflow conditions are required to approximate actual water year loads for a given site, constituent, and water year. Potential sources of these data were considered throughout the United States to evaluate load-estimation methods among multiple constituents and from sites with varied environmental settings and water-quality transport characteristics. Water-quality constituents evaluated include chloride, nitrate plus nitrite, total nitrogen, total phosphorus, and suspended sediment. Although it is desirable to evaluate loads of other types of constituents, such as pesticides and trace metals, long-term near-daily observations of these constituents are not available. Additionally, although specific conductance was used to evaluate the ability of methods to estimate decadal loads in Lee and others (2016), specific conductance was not used in this study because chloride is considered a better indicator of major ion transport in U.S. streams and rivers. The annual sum of the sampled, daily loads used to evaluate estimation methods tested herein are termed "observed loads."

Because few sites across the United States have observations every day of a given water year, the number of daily observations per water year for sites in this study range from 193 to 366, meaning that observed annual loads in this study represent about one-half to one full water year. Datasets selected for this study typically had less than 1 percent censored (that is, "below detection") values; the most censored values were observed for nitrate concentrations at the Sandusky River near Fremont, Ohio (04198000, hereafter referred to as "SAND" [3.8 percent]), and Rock Creek at Tiffin, Ohio (04197170, hereafter referred to as "ROCK" [3.7 percent]; fig. 1). Although nearly all the estimation methods considered can accommodate censored data, censored data were omitted from this study. The authors acknowledge the potential for bias and variability in observed concentrations and loads because of sampling and analytical procedures and the omission of censored data. However, the goal of this study is to characterize the accuracy of load estimates, and thus it is not necessary for observations to exactly represent loading conditions for a given site and water-quality constituent. Potential accuracy issues related to observed values, such as nonrepresentative sampling methods or the removal the censored data, do not hinder the ability to assess the accuracy of load-estimation methods considered herein.

The following sources of data are used to evaluate load-estimation methods. Heidelberg University (2005) has collected near-daily water-quality observations of water-quality constituents at sites in the upper Midwest since 1976. Chloride, total nitrogen, nitrate plus nitrite, and total phosphorus data were selected from Heidelberg University sites with at least 10 years of water quality data and at least 200 samples per year (table 1). Multiple observations were sometimes recorded on a single day; in these cases, one observation was randomly selected to represent that day for computations of observed and estimated loads.

The USGS National Water Information System (U.S. Geological Survey, 2017) was used to obtain daily streamflow, suspended-sediment, and subdaily and daily value nitrate data from continuous sensors. Suspended-sediment data were obtained from sites in a variety of environmental settings, with varying drainage areas, and with at least 10 years of continuous record from 1948 through 2014 (table 1). Continuous nitrate plus nitrite data were used in addition to Heidelberg University (2005) data to expand the environmental settings and drainage areas for which nutrient loads are evaluated (table 1; fig. 1). Because mean daily concentrations are not always published at sites with continuous nitrate plus nitrite sensors, mean daily nitrate plus nitrite values for this study are occasionally computed from available subdaily time-series data (typically recorded at 15- or 60-minute increments). Loads computed at USGS continuous nitrate plus nitrite sites are only considered within the "Evaluation among Sampling Sites" section in this report. All other comparisons involving nitrate plus nitrite used Heidelberg University (2005) datasets exclusively so that roughly equivalent sites and periods are compared among nitrate plus nitrite, chloride, total nitrogen, and total phosphorus estimates.

It is important to note that the length of observed records varies among sampling sites (table 1), and thus evaluations among water-quality constituents and sampling strategies in this report are more heavily weighted toward specific sites. The authors decided that considering as many data as possible would allow for a more complete evaluation of load-estimation methods.

Sampling Strategies

The frequency and hydrologic condition in which samples are collected depend upon the objectives and budget of the water-quality sampling program. An objective of this study is to evaluate the suitability of sampling strategies, including those used by the USGS NWQN, for annual load estimation. Load-estimation methods are evaluated by selecting sampling days from observed records using various

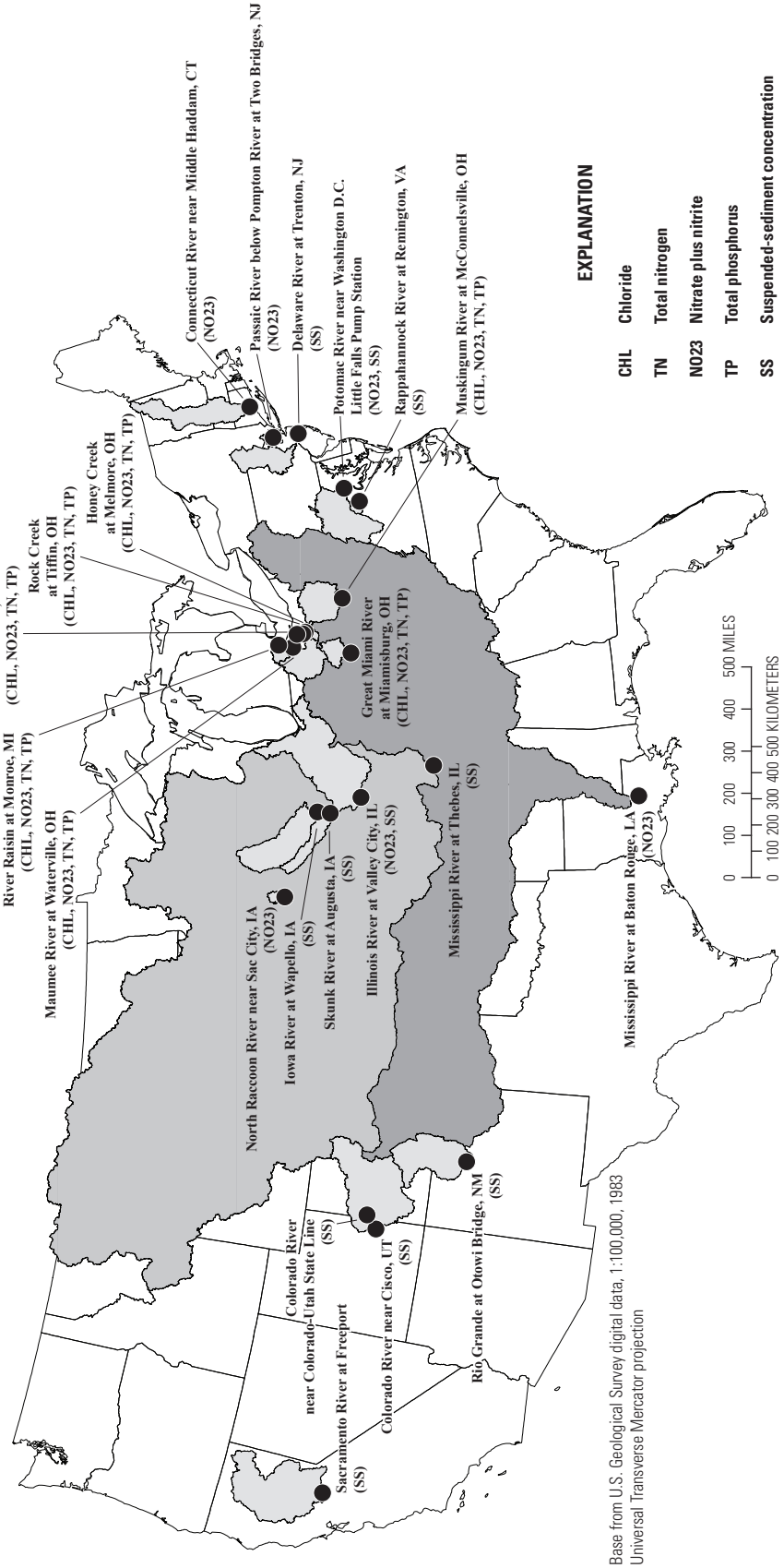


Figure 1. Locations of sites and basins used to evaluate load-estimation methods.

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

[USGS, U.S. Geological Survey; NJ, New Jersey; D.C., District of Columbia; VA, Virginia; OH, Ohio; IA, Iowa; IL, Illinois; NM, New Mexico; UT, Utah; CA, California; CT, Connecticut; MI, Michigan; LA, Louisiana]

USGS site name	Site abbreviation	USGS station number	Contributing drainage area (square miles)	Period of record	Percentage agriculture[1]	Percentage forest[1]	Percentage urban[1]	Coefficient of variation in daily streamflow
Suspended sediment (U.S. Geological Survey, 2017)								
Delaware River at Trenton, NJ	DELA	01463500	6,780	1949–82	15	67	10	1.00
Potomac River near Washington, D.C., Little Falls Pump Station	POTO	01646500	11,560	1960–91	30	59	10	1.42
Rappahannock River at Remington, VA	RAPP	01664000	619	1951–93	37	58	4	1.76
Maumee River at Waterville, OH	MAUM	04193500	6,330	1950–2003	79	6	11	1.62
Iowa River at Wapello, IA	IOWA	05465500	14,900	1978–2014	79	3	9	1.06
Skunk River at Augusta, IA	SKUN	05474000	4,312	1975–2014	77	7	7	1.52
Illinois River at Valley City, IL	VALL	05586100	26,743	1980–2011	70	10	15	0.77
Mississippi River at Thebes, IL	MISS	07022000	713,200	1982–2014	40	12	5	0.62
Rio Grande at Otowi Bridge, NM	RIO	08313000	14,300	1955–2014	4	42	1	1.01
Colorado River near Cisco, UT	COLO	09180500	24,100	1941–84	4	54	1	1.17
Sacramento River at Freeport, CA	SACR	11447650	27,233	1972–81	12	44	4	0.78
Chloride, total nitrogen, nitrate plus nitrite, and total phosphorus (Heidelberg, 2005, unless otherwise noted)								
Connecticut River at Middle Haddam, CT[2]	HADD	01193050	10,897	2013–16	8	70	12	0.78
Potomac River near Washington, D.C., Little Falls Pump Station[2]	POTO	01646500	11,560	2013–16	30	59	10	1.20
Muskingum River at McConnelsville, OH	MUSK	03150000	7,422	1995–2014	41	43	12	0.95
Great Miami River at Miamisburg, OH	GREA	03271601	2,715	1997–2014	72	9	17	1.35
River Raisin at Monroe, MI	RAIS	04176500	2,685	1983–2014	68	11	11	1.22
Maumee River at Waterville, OH	MAUM	04193500	6,330	1983–2014	79	6	11	1.17
Honey Creek at Melmore, OH	HONE	04197100	774	1977–2014	82	10	7	2.23
Rock Creek at Tiffin, OH	ROCK	04197170	35	1984–2014	79	11	9	3.18
Sandusky River near Fremont, OH	SAND	04198000	1,253	1983–2014 (no 2001)	81	9	8	1.87
North Raccoon River near Sac City, IA[2]	SAC	05482300	700	2009–16	89	0.3	7	1.60
Illinois River at Valley City, IL[2]	ILLI	05586100	26,743	2013–16	70	10	15	0.84
Mississippi River at Baton Rouge, LA[2]	BATO	07374000	1,125,810	2013–16	38	21	6	0.49

[1]Based on 2006 National Land Cover Database and aggregated from Falcone (2011) and the USGS Sediment Portal (Lee, 2013).

[2]Evaluated for nitrate only, data obtained from USGS National Water Information System (U.S. Geological Survey, 2017).

strategies; estimating loads using data from sampled days; and then comparing load estimates to the sum of observed daily loads for a given site, constituent, and water year. A total of 10 replicate datasets were randomly selected for each sampling site, constituent, and water year under the guidelines of each sampling strategy. Evaluating multiple replicates for each sampling site, constituent, sampling scenario, water year, and estimation facilitates a more robust evaluation of load-estimation methods. A total of six sampling strategies were evaluated in this report:

National Water Quality Network Sampling

The NWQN sampling strategy is included to evaluate USGS NWQN sampling procedures. In this strategy, one sample is taken from the observed record per month (selected randomly from days at least 24 days from the previous sample), and six additional samples are taken during months that typically have increased streamflow (and thus, loading) conditions. High-flow months are chosen at USGS NWQN sites based on seasonal patterns in rainfall and runoff; these high-flow months are typically the same at sites in similar geographic settings. In this study, high-flow months were determined by mimicking the sampling schedule at the nearest large inland river or coastal USGS NWQN site (Deacon and others, 2015; Lee and others, 2017a). When three samples were identified to be collected in a month, samples were required to be at least 7 days from the previous sample; when two samples were collected per month, samples were required to be at least 10 days from previous samples.

High-Flow Sampling

The high-flow sampling (HIFLOW) strategy is included to test if specifically targeting high streamflow conditions for water-quality sampling improves the accuracy of annual load estimates. It is important to note that this strategy benefits from prior knowledge of the timing and degree to which high streamflows occur and, thus, is an idealized scenario that could not be replicated in practice. The HIFLOW strategy is implemented by taking one sample per month (as with the NWQN strategy, samples are selected randomly at least 24 days from the previous sample), and an additional six samples are randomly taken from streamflows greater than the 80th percentile for the given water year. These high-flow samples are required to be 7 or more days from the previous water-quality sample.

High-Flow Early Sampling

The high-flow early sampling (HIFLOWE) strategy is designed to provide a more realistic evaluation of targeting high streamflows for sampling than the HIFLOW strategy. The USGS National Stream Quality Accounting Network program targeted high streamflows for sampling before 2006; however, concerns about missing high-flow periods often caused sampling crews to collect samples during the first observed high-flow events. This strategy resulted in sampling budgets frequently being spent before high-flow events that might have occurred later in the water year (C. Crawford, written commun., 2017). For the HIFLOWE strategy, one sample is taken randomly per month (at least 24 days from the previous sample) and six additional high-flow samples are taken during the first observed high streamflows (still defined as streamflows greater than the 80th percentile for a given water year) under the stipulation that samples are at least 7 days from any previous sample.

Biweekly Sampling

The biweekly sampling (BIWEEK) strategy is included to test the accuracy of load estimates obtained under relatively frequent sampling but without a specific emphasis on high streamflow conditions. For this strategy, samples are taken about once every 2 weeks by randomly selecting observations 12 to 16 days from the previous sample. This strategy represents the most frequent sampling of any tested herein.

Monthly Sampling

The monthly sampling (MONTH) strategy is designed to evaluate the effects of fixed-increment sampling under a reduced sampling frequency. The strategy takes one sample at random per month from the observed record while ensuring that samples are collected at least 24 days from the previous sample.

Bimonthly Sampling

The bimonthly sampling (BIMONTH) strategy is included to test the effects of infrequent sample collection on load-estimate accuracy. A total of six samples are taken at random from the observed record per year while requiring that samples are taken at least 54 days from the previous sample. This strategy represents the least frequent sampling of any tested herein.

Load-Estimation Methods

The following sections describe load estimation methods evaluated in this study. Estimation methods range from relatively simple approaches, such as similar linear interpolation, to relatively complex weighted regression methods. With the exception of the Weighted Regressions on Time, Discharge, and Season Method with Kalman Filtering (WRTDS_K), most of the methods considered in this study are similar to those described in an evaluation of methods for computing decadal loads (Lee and others, 2016).

Interpolation

The method of interpolation among subsequent water-quality samples (INTERP; table 2) represents the simplest of all estimation models considered. The INTERP method estimates daily concentration values by linearly interpolating over the set of sampled concentration values. The time series of interpolated, daily concentration values are then multiplied by daily streamflows and a conversion factor to obtain daily loads, which are then summed for the target water year. The INTERP method is implemented using the loadflex package (Appling and others, 2015) through the R statistical platform (R Core Team, 2017).

Beale's Ratio Estimator

The Beale's ratio estimator has been described widely (Beale, 1962; Tin, 1965; Dolan and others, 1981) and has been used primarily for load computation at sites contributing to the Great Lakes. Ratio estimators are typically implemented by delineating different strata within the sampled record. Strata may be defined based on time or streamflow conditions. Once strata are selected, Beale's estimator for the ratio of a given stratum is given by

Table 2. Estimation methods considered.

Estimation method abbreviation	Estimation method description	Number of years of data considered
INTERP	Interpolation of sampled values.	1
RATIO_T	Beale's ratio estimation with time-based stratification.	1
RATIO_F1	Beale's ratio estimation with flow-based stratification.	1
RATIO_F5	Beale's ratio estimation with flow-based stratification on the most recent 5 years of data.	5
L1	LOADEST stock 1-parameter model with streamflow as the only explanatory variable.	1
L5	LOADEST stock 5-parameter model with streamflow, season, and time as explanatory variables.	5
L7	LOADEST stock 7-parameter model with streamflow, streamflow squared, season, time, and time squared as explanatory variables.	5
LAICO	LOADEST stock "best selection" model that selects explanatory variables with the minimum Akaike information criteria.	5
AIC	LOADEST minimum Akaike information criteria method with additional explanatory variables as described in Lee and others (2017a).	5
PVAL	LOADEST minimum probability-value method with additional explanatory variables as described in Lee and others (2017a).	5
AIC_COMP	AIC method with an additional adjustment of daily estimates by the composite method (Aulenbach and Hooper, 2006).	5
WRTDS	Weighted Regressions on Time, Discharge, and Season method (Hirsch and others, 2010).	14 or more years[1]
WRTDS_K	WRTDS method with an additional adjustment of daily estimates by a Kalman filter methodology.	14 or more years[1]

[1]If the requisite number of samples is available, 14 years of observations are used; otherwise additional years are considered until 100 samples (or the 90 percent of the number of samples if less than 100 samples are available over the entire record) are reached.

$$\hat{R}\left(\frac{1+\frac{1-f}{n}c_{LQ}}{1+\frac{1-f}{n}c_{QQ}}\right) \quad (1)$$

where

$\hat{R} = \overline{l} / \overline{q}$ is the ratio of the stratum sample means of load, $\overline{l}$, and streamflow, $\overline{q}$;

$f = n / N$ is the ratio of the number of sampled days in the stratum, n, to the total number of days (sampled and unsampled) in the prediction period occurring in the stratum, N;

$c_{LQ} = s_{LQ} / \left(\overline{l}\overline{q}\right)$ is the ratio of the stratum sample covariance between load and streamflow, s_{LQ}, to the product of the stratum sample means of load and streamflow; and

$c_{QQ} = s_Q^2 / \overline{q}^2$ is the ratio of the stratum sample variance of streamflow to the square of the stratum sample mean of streamflow.

The estimate of load for all days within a given stratum is the sum of daily load in the sample plus the product of Beale's ratio estimate for the stratum, multiplied by the total streamflow for all unsampled days in the stratum. The summation of these estimates across all strata provides the total load estimate for all days in the prediction period. Beale's ratio estimates typically use discrete sample and streamflow data exclusively from the year in which loads are being computed (in contrast to most other methods evaluated herein). The most recent (2018) documented use of the Beale's ratio estimator (Maccoux and others, 2016) computed annual phosphorus loads from streams contributing to the Great Lakes using time-based strata chosen by a water-quality analyst.

In this study, three ratio estimators (table 2) that define strata in different ways are evaluated. The Beale's ratio estimation with time-based stratification (RATIO_T) method is implemented using the AutoBeale FORTRAN program (available on the USGS Github website at https://github.com/smwesten-usgs/AutoBeale), a commonly used iteration of the ratio estimator published originally by Richards (1998). This method uses water-quality and streamflow data from the target year only. As many as four strata are selected by date under this method; the number and timing of strata are defined to minimize the mean-squared error (MSE; computed using methods described in Baun [1982]) of the estimate. The MSE of the estimate is minimized by first computing the MSE for all possible dates with one stratum and selecting the stratum date that produces the smallest MSE. Successive strata are then tested and chosen contingent on the location of the first stratum until the number and locations of the strata result in the smallest possible MSE. The program then uses an adjustment procedure in which each stratum is tested at all possible dates between other strata until the MSE is minimized or has improved by less than 0.5 percent. See https://github.com/smwesten-usgs/AutoBeale/blob/master/doc/AUTOBEAL.pdf for more details. The number and dates of the strata are defined separately for each site, sampling strategy, water year, and replicate.

Other ratio estimators considered in this study define strata based on streamflow conditions and are selected to minimize the total MSE of the estimate. The Beale's ratio estimation with streamflow-based stratification (RATIO_F1) method uses data from the target year only, providing a useful comparison to the RATIO_T method. The Beale's ratio estimation with streamflow-based stratification on the most recent 5 years of data (RATIO_F5) method uses samples from the target water year and the 4 years before the target water year, making it more comparable to regression-based methods described later. The number and locations of strata were selected to minimize the MSE using a minimization routine implemented using the mgcv genetic algorithm package (Wood, 2006) in R (R Core Team, 2017). A maximum of 2 strata were used for the RATIO_F1 method; the consideration of additional samples allowed a maximum of 9 strata to be evaluated for the RATIO_F5 method. As with the RATIO_T method, the number and locations of streamflow strata for the RATIO_F1 and RATIO_F5 methods were chosen separately for every site, sampling strategy, water year, and replicate.

LOADEST Methods

The USGS LOADEST program uses maximum likelihood estimation to develop regression relations that relate infrequently available concentration data to various explanatory variables derived from daily streamflow and decimal time. LOADEST assumes that model residuals are normally distributed with a constant variance (Runkel and others, 2004) and uses a minimum variance unbiased estimate of instantaneous load to correct for retransformation bias (Cohn and others, 1989). The accuracy of the retransformation bias corrections is particularly susceptible to the misspecification of the model (that is, failure to properly model curvature in the relation and [or] heteroscedastic errors).

LOADEST model forms evaluated in this study are listed below (abbreviations for models are shown in parentheses). With the exception of the LOADEST with streamflow only method (L1; table 2), regression-based methods use sample data from a 5-year moving window (that is, loads are estimated from data obtained during the target year and the preceding 4 years). This is the same approach used for load estimation at USGS NWQN sites.

- LOADEST stock 1-parameter model with streamflow as the only explanatory variable (L1), using only data from the target year

$$\ln(C_t)=\beta_1+\beta_2 \ln Q_t+e_t \quad (2)$$

where

$\ln(C_t)$ is the natural logarithm of the constituent concentration for period t, assumed to be a day;

βk, k=1,…,2 are the model parameters to be estimated;
$\ln(Q_t)$ is the natural logarithm of mean daily discharge; and
e_t is the model residual.

- LOADEST stock 5-parameter model with streamflow, season, and time as explanatory variables (L5)

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

where

T_t is decimal time.

- LOADEST stock 7-parameter model with streamflow, streamflow squared, season, time, and time squared as explanatory variables (L7)

$$\ln(C_t)=\beta_1+\beta_2 \ln Q_t+\beta_3 \ln(Q_t)^2+\beta_4 T_t+\beta_5 T_t^2 +\beta_6 \sin(2\pi T_t)+\beta_7 \cos(2\pi T_t)+e_t \quad (4)$$

- LOADEST stock "best selection" model that selects explanatory variables with the minimum Akaike information criteria (LAICO)

This method considers all 11 original LOADEST model forms (see Runkel and others, 2004) and selects the regression equation that results in the smallest Akaike information criteria (Akaike, 1974) value. In its most complex form, the model form is that of the L7 method. This method is included in the LOADEST software package (Runkel and others, 2004).

- LOADEST minimum Akaike information criteria method with additional explanatory variables (AIC)

This method is similar to LAICO in that the model is selected among a population of models based on the minimum Akaike information criteria value; however, for this method, the list of possible explanatory variables is expanded from stock LOADEST options to include the logarithm of cubic streamflow and four variables indicative of historical streamflow conditions (Ryberg and Vecchia, 2012). These four variables are called "flow anomaly" variables. They are designed to capture the degree to which the discharge over some antecedent period departed from average conditions over multiple decades. Each streamflow anomaly variable is computed over a different antecedent period. The model with the minimum Akaike information criteria is selected from all possible combinations of all explanatory variables with the stipulation that the logarithm of streamflow is included as an explanatory variable. The four streamflow anomaly variables considered were adapted from Ryberg and Vecchia (2012) based on recommendations from Vecchia (written commun., 2014). These variables are defined as

$$FA\_1\_10\_DAY=X(t)-X_{10}(t) \quad (5)$$

$$FA\_1\_30\_DAY=X(t)-X_{30}(t) \quad (6)$$

$$FA\_30\_365\_DAY=X_{30}(t)-X_{365}(t) \quad (7)$$

$$FA\_100\_365\_ALL=(X_{100}(t)-X*(t))-(X_{365}(t)-X*) \quad (8)$$

where

$X(t)$ is the natural logarithm of mean daily discharge for day t,
$X_{10}(t)$ is the average of the natural logarithm of mean daily discharge for the 10 days up to and including day t,
$X_{30}(t)$ is the average of the natural logarithm of mean daily discharge for the 30 days up to and including day t,
$X_{365}(t)$ is the average of the natural logarithm of mean daily discharge for the 365 days up to and including day t,
$X_{100}(t)$ is the average of the natural logarithm of mean daily discharge for the 100 days up to and including day t, and
$X*(t)$ is the average of the natural logarithm of mean daily discharge for the period of record including day t.

- LOADEST minimum probability (p) values method with additional explanatory variables (PVAL)

This method is identical to AIC except that the minimum overall p-value is used to select the model from all potential combinations of explanatory variables (in contrast to the minimum Akaike information criteria).

- LOADEST minimum Akaike information criteria method with additional explanatory variables and adjustment via the composite method (AIC_COMP)

The AIC_COMP method is used to evaluate if adjusting daily estimates based on departures from sampled values improves the accuracy of annual load estimates. The composite method (Aulenbach and Hooper, 2006) was determined to improve the accuracy of decadal-load estimates relative to standard LOADEST estimates (Lee and others, 2016). This method is implemented by first computing the logarithm of daily estimated water-quality concentrations via the AIC method as described above. For this study, a linear interpolation is completed among modeled residuals (in logarithmic space); these interpolated values are then added to the original estimated daily values. Then, these values are retransformed, biased-corrected (using the same methods as in LOADEST), and multiplied by streamflow and a unit conversion to produce daily load estimates. The composite method part of this method is implemented using default options defined in the loadflex package (Appling and others, 2015) through the R statistical platform (R Core Team, 2017). Because Lee and others (2016) determined the magnitude of improvements among the composite method and FLUXMASTER (used in the development

of USGS SPAtially Referenced Regression On Watershed attributes [SPARROW] models) methods relative to standard LOADEST to be similar, the FLUXMASTER method is not evaluated in this study.

Weighted Regressions on Time, Discharge, and Season

The WRTDS method is implemented through the R package Exploration and Graphics for RivEr Trends (EGRET) (Hirsch and others, 2015). WRTDS is used to develop a time-varying linear relation 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 and others, 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 uses a bias correction factor specific to each year, day, and discharge to adjust for retransformation bias (see Moyer and others, 2012; Hirsch and others, 2015). With one exception, WRTDS model estimates computed in this study use default values specified in the EGRET software, including a windowY setting of 7, a windowQ setting of 2, a windowS setting of 0.5, and an edgeAdjust setting of "true." An exception to the use of the default setting is the minimum number of observations setting (minNumObs), which was changed to the lowest of 90 percent of the sample size or 100 (the default is that it is always 100) to facilitate model estimates for datasets with smaller sample sizes. Although WRTDS is designed to use data after the target year, the WRTDS and WRTDS_K models are estimated only after at least 5 years of data are collected and only consider data from the target year and years before the target year.

The WRTDS_K method is identical to WRTDS but includes an adjustment of the daily load estimates based on the observed residuals in logarithmic space. The concept is a simple approximation of the idea of a Kalman filter (hence the abbreviation WRTDS_K; Kalman, 1960). On days with water-quality observations, WRTDS_K uses observed values instead of the daily estimates produced by the WRTDS model. On days without observed values, residuals are generated using an autoregressive lag function and added to the daily estimates generated from WRTDS. For each set of intervening days without observations, a set of residual values is computed by a Monte Carlo simulation that is conditioned by the observed residuals on each end of the unsampled interval. These generated residuals have an autoregressive lag-1 structure with a serial correlation coefficient of 0.95. These residuals are then added to the expected value of the logarithm of concentration determined by the WRTDS model for that day. The logarithmic concentration values in this set are exponentiated to form a series of concentration values. A total of 50 replicates of this Monte Carlo simulation are completed. The WRTDS_K estimate for each of these intervening days is the mean of the 50 replicate values for that day. Further details on this method are presented in appendix 1.

In general, when the sampling is sparse, such as more than 60 days between observed values, the WRTDS_K estimates near the middle of those gaps will be similar to those determined in the original WRTDS method. For days near the samples, the WRTDS_K estimates will be quite different from the standard WRTDS estimates because they are strongly affected by the sampled concentrations. When there are only a few days between observations (less than or equal to 7 days), the WRTDS_K approach can produce estimates that are quite different from those determined in the original WRTDS method because measured data values will be used instead of standard WRTDS estimates, and the serial dependence of these data is likely have a strong effect on the estimates. The assumption that the autoregressive lag-1 correlation coefficient is 0.95 is consistent with experience with high frequency sampling data. Further research is being completed to attempt to optimize the selection of this coefficient, but preliminary results indicate that using 0.95 leads to results that are reasonably good, even if not optimal.

Evaluation of Load-Estimation Methods

Load-estimation methods are evaluated by the percentage difference of estimated annual loads from observed annual loads. This percentage is computed as

$$PercDiff = 100 * \frac{EST_{ijkl} - OBS_{jkl}}{OBS_{jkl}} \quad (9)$$

where

$PercDiff$ is the difference of estimated loads from observed loads, in percent;

EST_{ijkl} is the estimated annual load for sampling strategy i, water-quality constituent j, water year k, and sampling site l; and

OBS_{jkl} is the observed annual load for water-quality constituent j, water year k, and sampling site l.

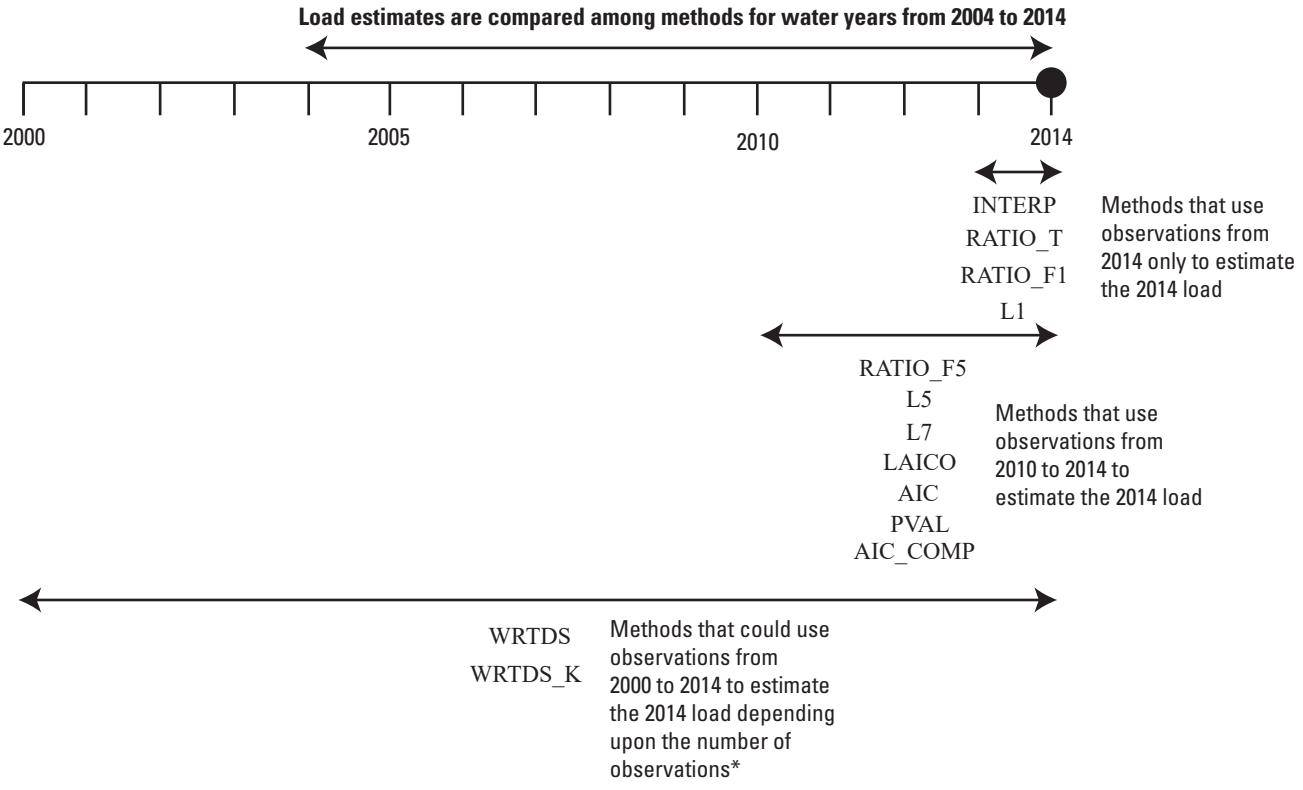
Any negative loads or loads greater than 10,000 times the observed loads were removed from consideration to facilitate the presentation of results and because analysts would likely be able to identify these estimates as erroneous in practice. These instances were relatively rare; negative loads occurred in 0.01 percent of cases, whereas estimates greater than 10,000 times the observed loads occurred in 0.03 percent of cases. Negative loads and loads with extreme positive bias primarily occurred at specific sites with the smallest drainages and variable streamflow conditions. ROCK recorded the most negative loads for total nitrogen, nitrate plus nitrite, and total phosphorus estimates (0.1 percent of possible ROCK total nitrogen, nitrate plus nitrite, and total phosphorus estimates),

whereas the Rappahannock River at Remington, Virginia, site (USGS station 01664000, hereafter referred to as "RAPP") recorded the most loads with extreme positive bias (0.9 percent of possible RAPP estimates).

For each sampling strategy, water-quality constituent, and sampling site, individual water year estimates are summarized by the percentage of annual estimates that fall within predefined threshold percentages (typically plus or minus [±] 20 percent) of the observed load. These thresholds are used as qualitative measures of the "acceptable" error of an annual load estimate to simplify the presentation of results; however, because different applications have different accuracy requirements, boxplots of estimation method errors are provided in appendixes 3 and 4. Data analyzed during this study are available as a USGS data release (Lee, 2019). The data release includes water-quality concentrations and daily streamflow data used to compute observed annual loads, observed annual loads computed from these data, and annual load estimates.

When possible, comparisons of estimation method accuracy are done using the same sites, years, and water-quality constituents. The way in which estimation methods would use water-quality observations to compute loads for a hypothetical record of data collected from 2000 to 2014 is illustrated in figure 2. In this example, most methods (RATIO_F5, L5, L7, LAICO, AIC, PVAL, and AIC_COMP) would use a backward looking, 5-year sampling window and thus would use data from 2010 to 2014 to estimate the load in 2014 (fig. 2). The INTERP, RATIO_T, RATIO_F1, and L1 methods would use data for 2014 only to estimate loads in 2014, whereas the WRTDS and WRTDS_K methods would use data from at least 2001–14 to estimate loads in 2014 (fig. 2). In this example, all methods could generate annual load estimates from 2004 to 2014, but only the methods that do not use historical data (INTERP, RATIO_T, RATIO_F1, and L1) also could produce estimates for 2000–3. Methods that use historical data (RATIO_F5, LOADEST-based methods, WRTDS, and WRTDS_K) require 5 years of data (4 years of historical data), so they would not produce annual estimates for the 2000–3 period. Thus, evaluations of estimation method performance in this scenario would only consider load estimates from 2004 to 2014 to ensure that equivalent records are compared among methods.

Estimation methods may perform differently depending on the amount of historical water-quality observations considered (fig. 2). Because many methods have the capacity to use more or less historical (or future) data, an additional evaluation of estimation method accuracy among different "sampling windows" is included in appendix 5. In this study,



*In this study, Weighted Regressions on Time, Discharge, and Season (WRTDS) and the WRTDS method with Kalman filtering (WRTDS_K) considered data from the target year and prior years only. WRTDS and WRTDS_K were implemented to consider 14 years of data if at least 100 observations were collected in those years. However, if fewer than 100 observations were present in the 14-year window, WRTDS and WRTDS_K would expand beyond 14 years until the minimum number of observations is met. In this study, the minumum number of observations is set to the smaller of 100 or 90 percent of the number of samples in the dataset.

Figure 2. Schematic illustration of data used by methods to estimate loads for a hypothetical sampling record from 2000 to 2014.

WRTDS and WRTDS_K are implemented to use data from the target year and previous 13 years when available (the default windowY setting of 7; see Hirsch and others [2015] for more details). However, in contrast to other methods considered herein, (1) the effect of historical water-quality observations on WRTDS and WRTDS_K estimates varies depending upon streamflow, season, and when the sample was collected, and (2) observations may be used beyond the most recent 14 years depending upon the "minimum number of observations" setting in WRTDS and WRTDS_K (the minNumObs argument; see Hirsch and others [2015] for more details). Thus, in the example in figure 2, if more than 110 observations were collected across the period of record (2000–14) but fewer than 100 observations were recorded from 2001 to 2014, WRTDS and WRTDS_K would add additional data before 2001 (starting in 2000 and going backward) until 100 observations were reached. However, if only 90 observations were collected across the period of record (2000–14), WRTDS and WRTDS_K would look backward from 2001 until 81 observations were reached (90 percent of 90 observations). Although WRTDS (and other methods) has the capacity to use data after the target year to compute loads, only data from the target year and before the target year are considered in this study.

Results of Method Performance Evaluations

Practitioners are commonly required to compute loads using data from ambient monitoring networks in which samples are collected based on multiple objectives and are subject to funding limitations. In tables 3–7, a frame of reference is provided regarding the approximate level of accuracy expected when computing chloride, total nitrogen, nitrate plus nitrite, total phosphorus, and suspended-sediment loads under a given sampling strategy and estimation method, although it is important to note that results are specific to the sampling sites and periods evaluated in this study. Estimates from sites that were evaluated for nitrate plus nitrite only (table 1) were omitted from tables 3–7 to facilitate comparability among the various constituents. Estimation methods are sorted based on the percentage of estimates within ±20 percent of observed loads across all sampling strategies. An additional table of the percentage of estimates within ±10 percent of observed loads is included in appendix 2 for practitioners interested in an alternative measure of load-estimate accuracy.

Table 3. Percentage of annual chloride load estimates within plus or minus 20 percent of observed loads.

[Rows are sorted by methods with the most to least estimates within criteria in aggregate. Data are shaded in ranges of 90–100 percent, 80–89 percent, 70–79 percent, 60–69 percent, 50–59 percent, 40–49 percent, and 30–39 percent. BIWEEK, biweekly sampling; HIFLOW, high-flow sampling; HIFLOWE, high-flow early sampling; NWQN, National Water Quality Network sampling; MONTH, monthly sampling; BIMONTH, bimonthly sampling]

Method	BIWEEK, in percent	HIFLOW, in percent	HIFLOWE, in percent	NWQN, in percent	MONTH, in percent	BIMONTH, in percent
WRTDS_K	99	98	98	96	96	91
AIC_COMP	99	97	98	96	95	84
AIC	99	97	97	95	93	83
PVAL	98	97	97	94	93	83
L7	97	95	96	93	94	84
LAICO	96	94	95	93	92	83
WRTDS	92	92	92	90	90	84
L5	89	91	90	86	88	83
L1	89	87	81	87	80	75
RATIO_F5	83	80	80	79	77	72
RATIO_F1	82	73	72	72	63	51
RATIO_T	79	72	71	71	66	51
INTERP	63	57	56	51	49	39

Table 4. Percentage of annual total nitrogen load estimates within plus or minus 20 percent of observed loads.

[Rows are sorted by methods with the most to least estimates within criteria in aggregate. Data are shaded in ranges of 90–100 percent, 80–89 percent, 70–79 percent, 60–69 percent, 50–59 percent, and 40–49 percent. BIWEEK, biweekly sampling; HIFLOW, high-flow sampling; HIFLOWE, high-flow early sampling; NWQN, National Water Quality Network sampling; MONTH, monthly sampling; BIMONTH, bimonthly sampling]

Method	BIWEEK, in percent	HIFLOW, in percent	HIFLOWE, in percent	NWQN, in percent	MONTH, in percent	BIMONTH, in percent
WRTDS_K	96	93	92	87	89	72
AIC_COMP	92	86	85	84	80	60
RATIO_T	90	81	82	83	75	60
INTERP	88	80	81	80	74	63
RATIO_F1	86	81	83	77	74	60
AIC	82	78	77	72	75	57
WRTDS	76	74	72	72	71	65
PVAL	79	76	75	70	71	57
L7	75	75	75	67	74	62
LAICO	73	73	72	65	69	58
RATIO_F5	64	64	61	66	61	62
L1	51	50	50	43	51	45
L5	48	50	51	43	51	46

Table 5. Percentage of annual nitrate plus nitrite load estimates within plus or minus 20 percent of observed loads.

[Rows are sorted by methods with the most to least estimates within criteria in aggregate. Data are shaded in ranges of 90–100 percent, 80–89 percent, 70–79 percent, 60–69 percent, 50–59 percent, 40–49 percent, 30–39 percent, and 0–29 percent. BIWEEK, biweekly sampling; HIFLOW, high-flow sampling; HIFLOWE, high-flow early sampling; NWQN, National Water Quality Network sampling; MONTH, monthly sampling; BIMONTH, bimonthly sampling]

Method	BIWEEK, in percent	HIFLOW, in percent	HIFLOWE, in percent	NWQN, in percent	MONTH, in percent	BIMONTH, in percent
WRTDS_K	93	88	88	84	82	64
INTERP	93	85	86	86	78	69
RATIO_T	91	81	82	83	76	59
AIC_COMP	87	76	78	76	70	49
RATIO_F1	83	76	78	72	68	59
WRTDS	69	67	67	66	65	56
RATIO_F5	57	55	53	59	53	56
AIC	55	53	63	43	53	40
PVAL	54	52	61	42	52	39
L7	50	53	60	41	53	43
LAICO	50	51	59	42	41	41
L5	28	28	30	22	30	28
L1	26	28	27	24	30	31

Table 6. Percentage of annual total phosphorus load estimates within plus or minus 20 percent of observed loads.

[Rows are sorted by methods with the most to least estimates within criteria in aggregate. Data are shaded in ranges of 80–89 percent, 70–79 percent, 60–69 percent, 50–59 percent, 40–49 percent, 30–39 percent, and 0–29 percent. BIWEEK, biweekly sampling; HIFLOW, high-flow sampling; HIFLOWE, high-flow early sampling; NWQN, National Water Quality Network sampling; MONTH, monthly sampling; BIMONTH, bimonthly sampling]

Method	BIWEEK, in percent	HIFLOW, in percent	HIFLOWE, in percent	NWQN, in percent	MONTH, in percent	BIMONTH, in percent
WRTDS_K	85	76	75	73	69	58
WRTDS	79	72	71	71	66	57
RATIO_F5	76	69	69	72	66	53
AIC_COMP	75	67	69	66	60	46
AIC	71	67	67	66	60	48
PVAL	69	67	65	67	59	50
L5	68	63	59	65	61	50
L7	61	61	61	64	57	49
LAICO	62	61	59	63	57	49
L1	56	51	50	55	47	33
RATIO_F1	58	48	49	50	42	28
RATIO_T	48	37	37	39	35	28
INTERP	37	29	28	28	25	21

Table 7. Percentage of annual suspended-sediment load estimates within plus or minus 20 percent of observed loads.

[Rows are sorted by methods with the most to least estimates within criteria in aggregate. Data are shaded in ranges of 70–79 percent, 60–69 percent, 50–59 percent, 40–49 percent, 30–39 percent, and 0–29 percent. BIWEEK, biweekly sampling; HIFLOW, high-flow sampling; HIFLOWE, high-flow early sampling; NWQN, National Water Quality Network sampling; MONTH, monthly sampling; BIMONTH, bimonthly sampling]

Method	BIWEEK, in percent	HIFLOW, in percent	HIFLOWE, in percent	NWQN, in percent	MONTH, in percent	BIMONTH, in percent
WRTDS_K	70	71	69	64	54	44
AIC_COMP	70	69	68	62	52	36
AIC	56	60	56	54	48	36
PVAL	56	59	54	53	47	35
L7	53	56	50	52	47	36
LAICO	51	55	48	50	44	36
WRTDS	48	50	46	48	45	40
L1	51	55	50	50	42	30
INTERP	55	56	53	48	37	24
RATIO_F5	49	49	49	47	43	35
RATIO_T	53	52	50	46	36	25
RATIO_F1	54	40	40	46	37	25
L5	38	44	42	35	36	33

Although all estimation methods have the potential to produce accurate load estimates, selected methods are more likely to do so than others. When considering all sampling strategies in aggregate, WRTDS_K produced the most estimates within ±20 percent of observed loads among all water-quality constituents; AIC_COMP produced the second, third, or fourth most estimates within these thresholds for chloride, total nitrogen, nitrate plus nitrite, and suspended-sediment estimates; and WRTDS produced the second most estimates within ±20 percent of observed total phosphorus loads. Although the WRTDS_K and AIC_COMP methods were the most accurate generally, the performance of some methods, such as INTERP and ratio estimators, varied substantially for different water-quality constituents. INTERP, RATIO_F1, and RATIO_T were among the most accurate methods for computing total nitrogen and nitrate plus nitrite loads, whereas RATIO_F5 was one of the most accurate methods for computing total phosphorus loads. With a few exceptions, regression-based methods that considered cubic streamflow and streamflow anomaly variables (AIC and PVAL) produced more estimates within ±20 percent of observed loads than regression-based estimates that relied on linear or quadratic representations of concentration/streamflow relations (in logarithmic space) exclusively (L1, L5, L7, and LAICO). Further analysis of differences in method performance among sampling strategies, water-quality constituents, and sampling sites is detailed in the following sections.

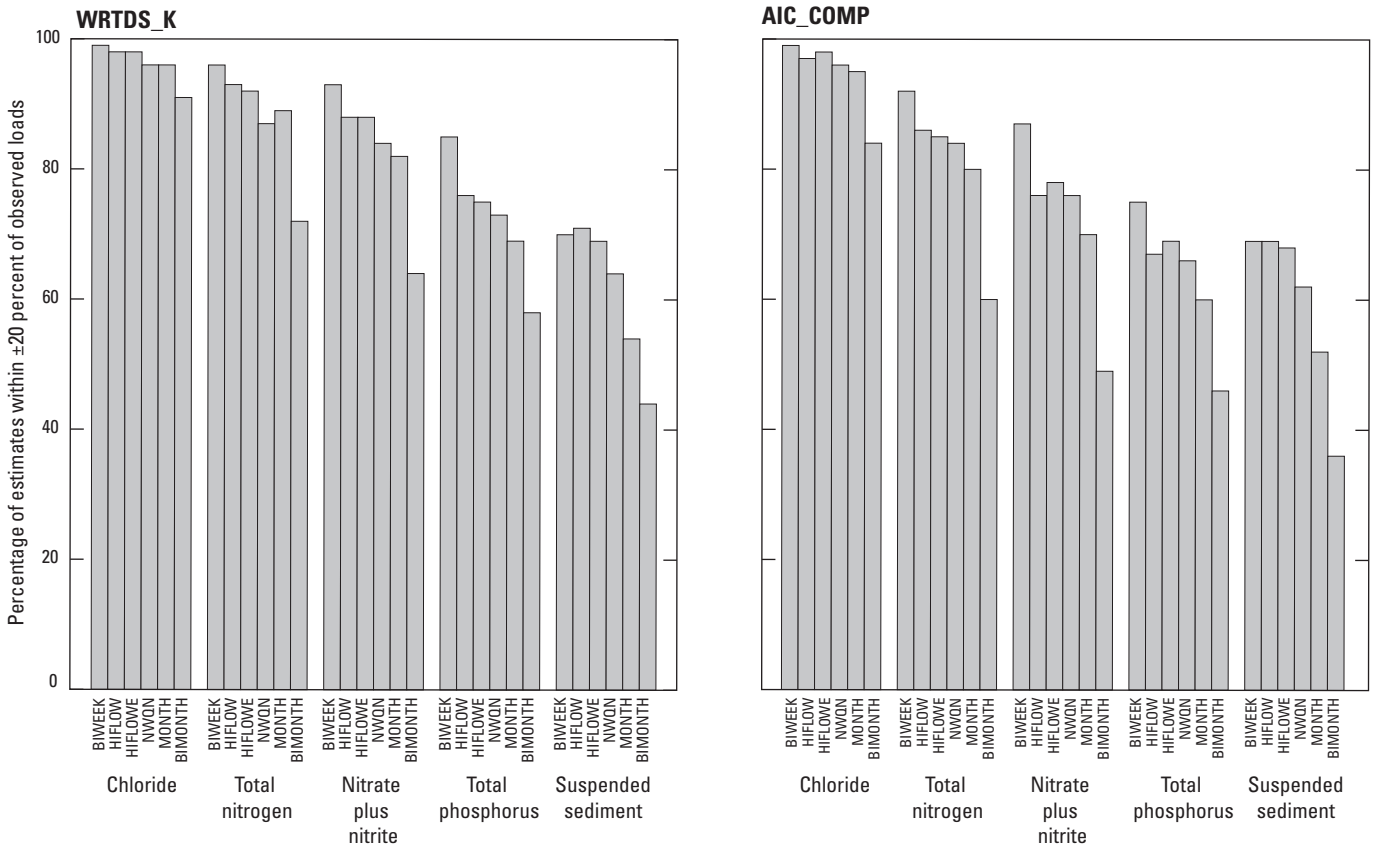
Evaluation of Sampling Strategies

Sampling strategies are compared to guide practitioners regarding the best network design for computing annual loads and to evaluate existing USGS NWQN procedures. The percentage of estimates within ±20 percent of observed loads is compared in figure 3 by sampling strategy for AIC_COMP and WRTDS_K, which were the best performing methods across water-quality constituents in tables 3–7. Plots in appendix 3 (figs. 3.1–3.13) show the distribution of errors for all estimation methods among sampling strategies and water-quality constituents. Sampling strategies in figure 3 and appendix 3 are ordered from left to right by decreasing numbers of samples per year. Comparisons of estimation method accuracy among strategies with different sampling frequencies allow practitioners to evaluate the degree to which additional water-quality sampling may or may not improve the accuracy of annual load estimates.

Strategies with more frequent sample collection and targeted high-flow sampling generally produced more accurate load estimates. Among all water-quality constituents and estimation methods, BIWEEK (69 percent) had slightly more estimates within ±20 percent of observed loads than strategies with 18 samples per year (HIFLOW, HIFLOWE, and NWQN; 64–67 percent) or 12 samples per year (MONTH, 61 percent). The BIMONTH (51 percent) strategy had the fewest samples within the ±20-percent threshold. Sampling strategies were similarly grouped in terms of the percentage of estimates with extreme errors, defined as those more than double (100 percent greater) or less than half (less than −50 percent) of observed loads. BIWEEK, HIFLOW, HIFLOWE, NWQN, and MONTH had similar percentages of estimates with extreme errors (23–24 percent), whereas BIMONTH (26 percent) had slightly more estimates outside of this threshold.

The effect of sampling strategy on load estimates differed somewhat among water-quality constituents and specific estimation methods (fig. 3 and figs. 3.1–3.13). WRTDS_K and AIC_COMP each produced between 95 and 99 percent of chloride estimates within ±20 percent of observed loads (fig. 3) using the BIWEEK, HIFLOW, HIFLOWE, NWQN, and MONTH sampling strategies but had fewer estimates within this threshold using the BIMONTH sampling strategy (WRTDS_K, 91 percent; AIC_COMP, 84 percent). WRTDS_K and AIC_COMP produced the most total nitrogen and nitrate plus nitrite estimates within ±20 percent of observed loads under the BIWEEK sampling strategy (fig. 3; tables 4–5). Among strategies with 18 samples per year, targeted high-flow sampling by the HIFLOW and HIFLOWE sampling strategies generally produced slightly more total nitrogen estimates within the ±20-percent threshold under WRTDS_K and AIC_COMP than under the NWQN strategy (fig. 3; tables 4–5). Reduced sampling under the MONTH strategy resulted in similar accuracy to the NWQN method for total nitrogen and nitrate plus nitrite loads using WRTDS_K but produced slightly fewer loads within ±20 percent of observed loads via the AIC_COMP method (fig. 4; tables 4–5). The BIMONTH sampling strategy resulted in substantially less accurate total nitrogen and nitrate plus nitrite loads with respect to the ±20-percent threshold under the WRTDS_K and AIC_COMP methods (fig. 3; tables 4–5).

The BIWEEK sampling strategy also improved the accuracy of WRTDS_K and AIC_COMP-computed total phosphorus loads relative to strategies with less frequent sampling (fig. 3). BIWEEK sampling through WRTDS_K (85 percent within ±20 percent of observed loads) and AIC_COMP (75 percent) was more accurate than total phosphorus loads computed via the HIFLOW, HIFLOWE, and NWQN strategies (73–76 percent for WRTDS_K and 66–69 percent for AIC_COMP). MONTH sampling resulted in slightly reduced accuracy (69 percent for WRTDS_K and 60 percent for AIC_COMP), whereas BIMONTH sampling substantially reduced estimation method accuracy relative to other sampling strategies (58 percent for WRTDS_K and 46 percent for AIC_COMP). In contrast to results observed for nitrate plus nitrite and total phosphorus, BIWEEK sampling offered little to no improvement in the accuracy of suspended-sediment estimates relative to strategies that targeted high streamflows with 18 samples per year. HIFLOW estimates produced by the WRTDS_K method resulted in the most estimates within ±20 percent of observed loads (71 percent), although the BIWEEK (70 percent) and HIFLOWE (69 percent) strategies demonstrated similar accuracy. AIC_COMP estimates were similar in accuracy to WRTDS_K; 68–70 percent of



EXPLANATION

[±, plus or minus]

WRTDS_K	**Weighted Regressions on Time, Discharge, and Season method with Kalman filtering**
AIC_COMP	**LOADEST minimum Akaike information criteria method with local adjustments for residual departures using the composite method**
BIWEEK	**Biweekly sampling**
HIFLOW	**High-flow sampling**
HIFLOWE	**High-flow early sampling**
NWQN	**National Water Quality Network sampling**
MONTH	**Monthly sampling**
BIMONTH	**Bimonthly sampling**

Figure 3. Percentage of estimates within plus or minus 20 percent of observed loads among water-quality constituents and sampling strategies for the WRTDS_K and AIC_COMP methods.

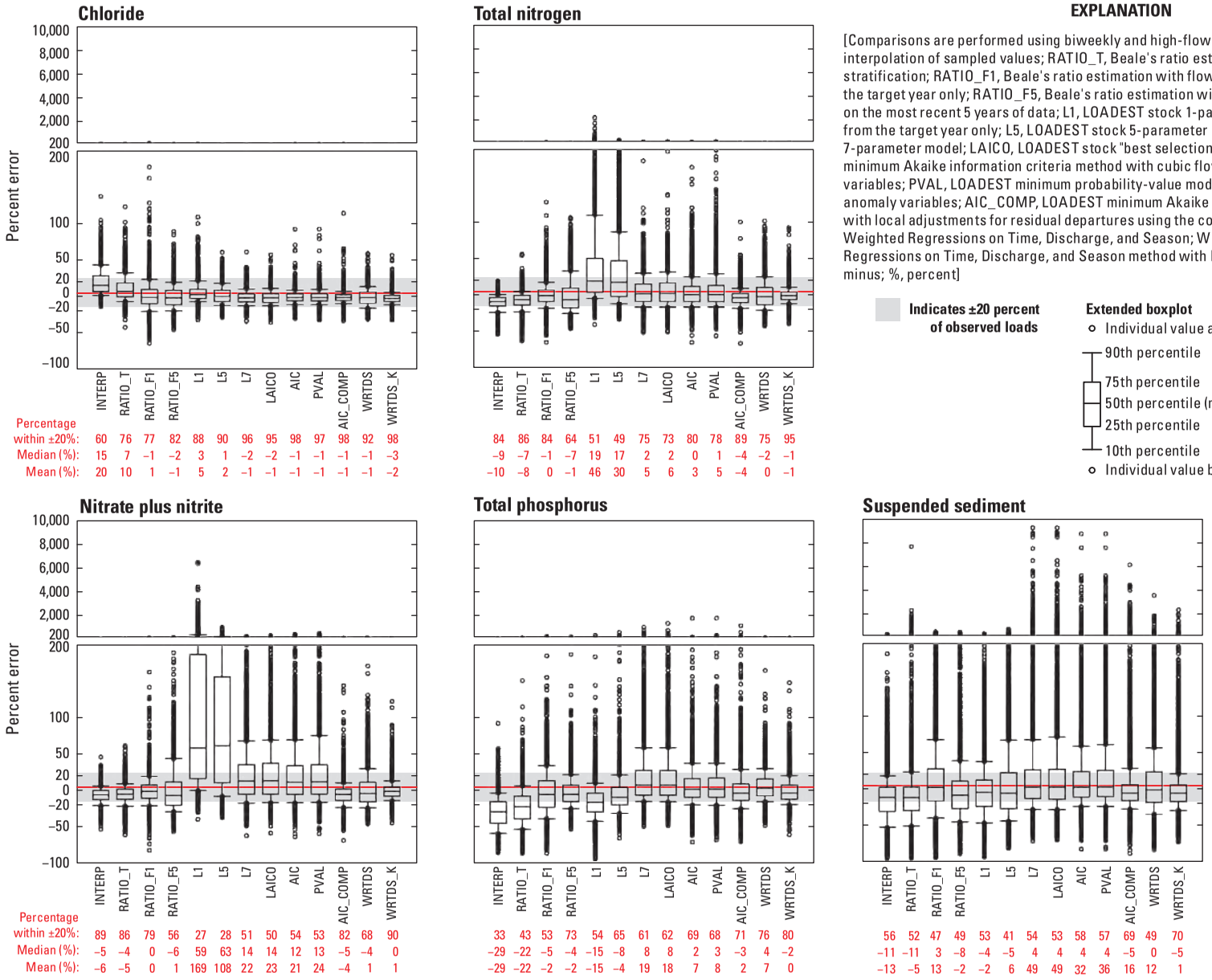


Figure 4. Comparison of estimation method accuracy among water-quality constituents.

estimates were within the ±20-percent threshold for BIWEEK, HIFLOW, and HIFLOWE strategies. Seasonal weighting of 18 samples per year via the NWQN strategy resulted in less accurate suspended-sediment loads (64 percent for WRTDS_K and 62 percent for AIC_COMP) as compared to the strategies that specifically targeted high-flow conditions. Further reductions in sample frequency via the MONTH (54 percent for WRTDS_K and 52 percent for AIC_COMP) and BIMONTH (44 percent for WRTDS_K and 36 percent for AIC_COMP) sampling strategies resulted in substantial further decreases in estimation method accuracy with respect to the ±20-percent threshold.

Although a single sampling strategy did not always produce the most accurate estimates for a given water-quality constituent, some general patterns were evident. The collection of only 6 samples per year substantially reduced the accuracy of estimates as compared to 12–26 samples, particularly when computing total nitrogen, nitrate plus nitrite, total phosphorus, and suspended-sediment loads. However, it is important to note that even under BIMONTH sampling, more than 70 percent of chloride estimates were still within ±20 percent of observed loads using most estimation methods, and 72 percent of BIMONTH total nitrogen estimates were within the ±20-percent threshold when using WRTDS_K. Increased sampling under the BIWEEK strategy (26 samples per year) offered moderate improvements in accuracy compared to strategies with frequencies of 18 samples per year. When 18 samples per year were collected, the purposeful collection of high-flow samples via the HIFLOW and HIFLOWE strategies generally produced more accurate load estimates than seasonally weighted (NWQN) sampling, although the degree of improvement varied across methods and constituents. The MONTH sampling strategy produced consistently fewer estimates within the ±20-percent threshold when compared to strategies with 18 samples per year that targeted high-flow conditions but still offered a substantial improvement in load-estimate accuracy as compared to the BIMONTH sampling strategy.

Evaluation of Methods among Constituents

Lee and others (2016) determined that the likelihood of computing accurate decadal water-quality loads varied substantially among water-quality constituents. This section evaluates the accuracy of estimation methods for computing annual loads among water-quality constituents using the HIFLOW and BIWEEK sampling strategies (which were generally determined to be the most accurate in the previous section). In general, estimation methods were the most accurate when computing chloride loads and were progressively less accurate when computing total nitrogen, nitrate plus nitrite, total phosphorus, and suspended-sediment loads.

Although chloride estimates were the most accurate (88 percent within ±20 percent of observed loads) among all estimation methods, regression and weighted-regression methods, including WRTDS_K, AIC_COMP, AIC, PVAL, L7, and LAICO, produced the most estimates within ±20 percent of observed loads (94–99 percent; table 3). Total nitrogen estimates were less accurate (76 percent of estimates within ±20 percent of observed loads) than chloride estimates generally and differed somewhat in terms of individual method accuracy. Methods that adjusted for departures from measured values (WRTDS_K, 95 percent; AIC_COMP, 89 percent) were the most accurate, whereas comparable methods that do not adjust daily estimates based on measured values (WRTDS and AIC) produced substantially fewer estimates within this threshold (75 percent and 80 percent, respectively). The L1 and L5 methods, which use linear relations among loads and streamflow (in logarithmic space), produced the fewest total nitrogen estimates within the ±20-percent threshold (49–51 percent) and tended to produce positively biased results (fig. 4).

Nitrate plus nitrite estimates were less accurate (63 percent of estimates across all methods within ±20 percent of observed loads) than total nitrogen estimates generally but demonstrated similar patterns among individual estimation methods. As with total nitrogen, methods that adjusted daily estimates based on departures from measured values (WRTDS_K, 90 percent; AIC_COMP, 82 percent) produced more estimates within ±20 percent of observed loads than estimates from methods without adjustments (WRTDS, 68 percent; AIC, 54 percent). Similarly, for total nitrogen, the L1 and L5 methods produced the fewest estimates within the ±20-percent threshold (27–28 percent) and tended to produce positively biased loads (fig. 4). However, in contrast to patterns observed with total nitrogen, INTERP (89 percent) produced the second most nitrate plus nitrite estimates within ±20 percent of observed loads, and regression-based methods other than L1 and L5 (L7, LAICO, AIC, and PVAL) that do not correct for departures from sampled values tended to produce positively biased nitrate plus nitrite loads (median +7–9 percent; mean +18–21 percent; fig. 4). Examples illustrating why selected LOADEST methods tended to produce biased total nitrogen and nitrate plus nitrite loads are included in the "Examples of Method Performance" section later in this report.

Total phosphorus estimates were less accurate than the previously described constituents (62 percent within ±20 percent of observed loads) and differed from total nitrogen and nitrate plus nitrite estimates in terms of individual method performance. In contrast to total nitrogen and nitrate plus nitrite estimates, methods that used data from the target year only (INTERP, RATIO_T, RATIO_F1, and L1) produced fewer total phosphorus estimates within the ±20-percent threshold (33–54 percent) in comparison to other methods. The WRTDS_K (80 percent) and WRTDS (76 percent) methods, which use weighted regression and more historical water-quality observations than other methods, as well as the RATIO_F5 (73 percent) method, produced the most estimates within ±20 percent of observed loads. Also in contrast to total nitrogen and nitrate plus nitrite results, the adjustment of daily

loads based on departures from observed values used in the WRTDS_K and AIC_COMP methods only resulted in slight improvements in accuracy compared to uncorrected methods (WRTDS and AIC; fig. 4). The consideration of cubic streamflow and streamflow anomalies in the AIC and PVAL methods produced slightly more estimates (68–69 percent) within the ±20-percent threshold than stock LOADEST methods L5, L7, and LAICO (61–65 percent).

Suspended-sediment estimates were the least accurate among water-quality constituents considered (55 percent within ±20 percent of observed loads among all methods). WRTDS_K and AIC_COMP methods produced the most estimates within the ±20-percent threshold (70 percent and 69 percent, respectively), substantially more than identical methods that do not adjust estimates based on measured values (WRTDS, 49 percent; AIC, 58 percent). RATIO estimators (RATIO_T, RATIO_F1, and RATIO_F5) and stock LOADEST methods (L1, L5, L7, and LAICO) were among the least accurate methods (41–54 percent of estimates within ±20 percent of observed loads) for computing suspended sediment. As with total phosphorus estimates, the consideration of cubic streamflow and streamflow anomaly terms via the AIC and PVAL methods produced more estimates within the ±20-percent threshold (57–58 percent) than stock LOADEST methods. Unlike total phosphorus estimates, the INTERP method (56 percent) produced more suspended-sediment estimates within the ±20-percent threshold than most estimation methods; however, many of these loads were biased low (fig. 4).

Evaluation among Sampling Sites

Lee and others (2016) determined that, for a given water-quality constituent, the accuracy of methods for estimating decadal loads varied substantially among sampling sites. To illustrate the importance of site-specific processes when estimating annual loads, method accuracy is compared among sampling sites and water-quality constituents. As in the previous section, differences in method performance are evaluated using only HIFLOW and BIWEEK sampling strategies.

For the WRTDS_K and AIC_COMP methods, which were generally the most accurate across multiple water-quality constituents (tables 3–7), the accuracy of annual load estimates among sampling sites had a clear, inverse relation with variability of observed daily loads (as measured by the coefficient of variation; fig. 5). Sites with more variable loading conditions typically have smaller drainages and more variable streamflow conditions. Selected sites with relatively few estimates within the ±20-percent threshold for a given constituent are indicated in figures 5*A* and *B*. ROCK had the most variable daily chloride, total nitrogen, nitrate plus nitrite, and total phosphorus loads and produced the fewest estimates within the ±20-percent threshold for each of these constituents (fig. 5*A, B*). A group of three sites, including RAPP, the Delaware River at Trenton, New Jersey, site (USGS station 01463500, hereafter referred to as "DELA"), and the Potomac River near Washington, D.C., Little Falls Pump Station site (USGS station 01646500, hereafter referred to as "POTO") had the most variable daily suspended-sediment loads and produced the fewest estimates within ±20 percent of observed loads. Although increased variability in daily loading conditions hindered estimation method performance generally, individual methods account for changing streamflow and water-quality concentrations differently, and thus the performance of specific estimation methods varied among sampling sites.

Chloride estimates were the least accurate at the three sites with the most variable observed loads, which include ROCK (74 percent within ±20 percent of observed loads); the Honey Creek at Melmore, Ohio, site (USGS station 04197100, hereafter referred to as "HONE"; 83 percent); and SAND (88 percent). These three sites together composed 82 percent of chloride estimates outside of the ±20-percent threshold (fig. 6; appendix 3). A total of 77 percent of chloride estimates outside of ±20 percent of observed loads were computed by the INTERP, RATIO_T, RATIO_F1, RATIO_F5, and L1 methods. As illustrated in previous sections, WRTDS_K and LOADEST-based methods produced the most accurate chloride estimates generally; these methods also produced the most estimates within ±20 percent of observed loads at the more variable ROCK, HONE, and SAND sites (fig. 6). As with chloride, total nitrogen estimates were the least accurate at sites with the most variable daily loads, which include ROCK (49 percent of estimates within ±20 percent of observed loads), HONE (73 percent), and SAND (74 percent). WRTDS_K and AIC_COMP methods generally produced the most estimates within the ±20-percent threshold at sampling sites, including the ROCK (81 percent and 67 percent, respectively), HONE (95 percent and 87 percent), and SAND (96 percent and 91 percent) sites.

As with previously described constituents, sites with more variable loading conditions (ROCK, 43 percent; SAND, 57 percent; HONE, 64 percent) produced among the fewest nitrate plus nitrite estimates within ±20 percent of observed loads. However, in contrast to patterns observed for chloride and total nitrogen loads, nitrate plus nitrite loads at the Maumee River at Waterville, Ohio (USGS station 04193500, hereafter referred to as "MAUM"), and the River Raisin at Monroe, Michigan (USGS station 04176500, hereafter referred to as "RAIS"), sites had comparatively more estimates outside of the ±20-percent threshold (57 and 71 percent, respectively), indicating that the nitrate plus nitrite transport regime at these sites made them difficult to represent. Among the sites in which nitrate plus nitrite loads were the least accurate (ROCK, MAUM, SAND, and HONE), the WRTDS_K, INTERP, and RATIO_T methods generally produced the most accurate load estimates. LOADEST-based methods that do not adjust records based on departures from measured values (L1, L5, L7, LAICO, AIC, and PVAL) produced the fewest nitrate plus nitrite estimates within the ±20-percent threshold at the ROCK, HONE, SAND (4–48 percent), and MAUM sites. In addition to sites monitored by Heidelberg

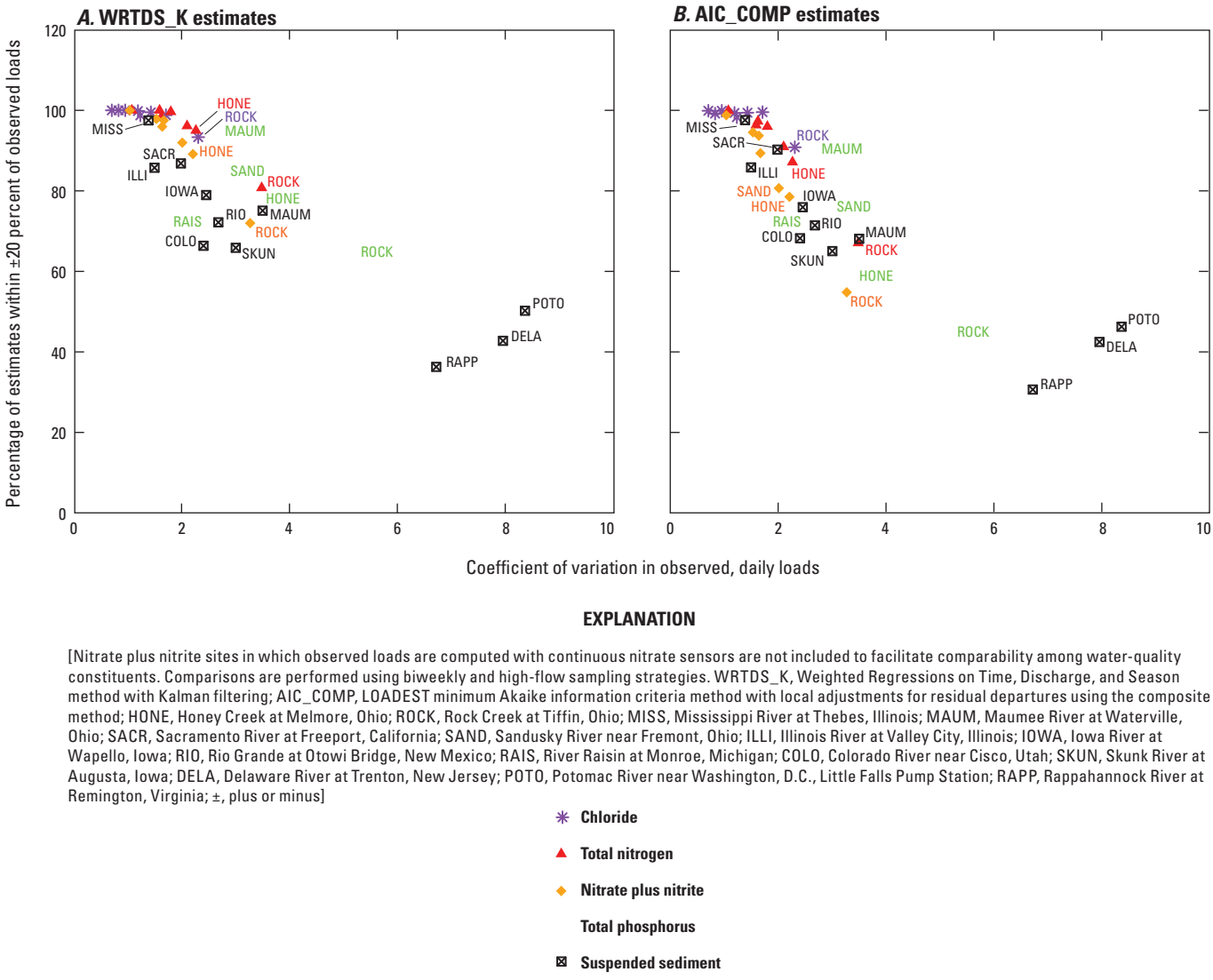
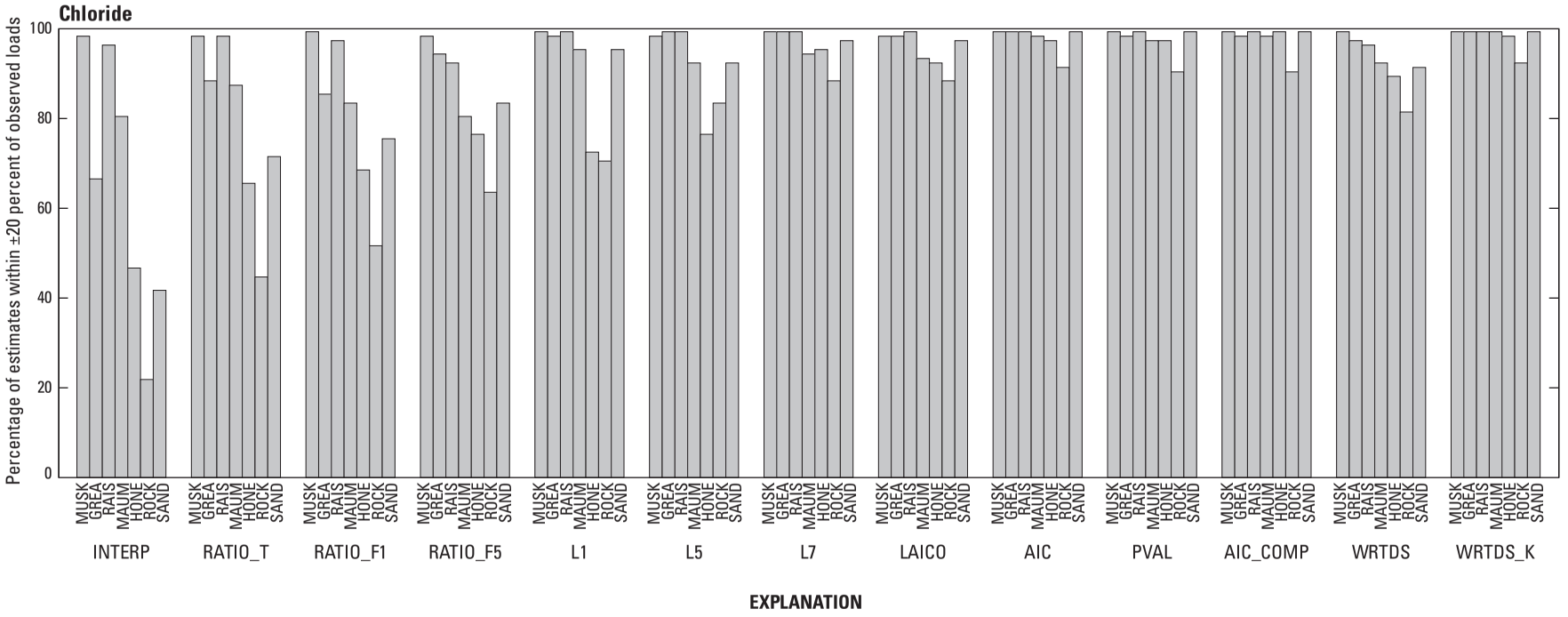


Figure 5. Percentage of load estimates within plus or minus 20 percent of observed loads compared to the variability of observed daily loads. *A*, WRTDS_K load estimates; and *B*, AIC_COMP load estimates.

University, nitrate plus nitrite loads also were computed at USGS sites equipped with continuous nitrate sensors to expand the range of drainage areas and environmental settings considered (table 1). These sites have relatively short record lengths (4–6 years), and thus methods were evaluated using a 3-year (as opposed to a 5-year) window to allow loads to be computed for multiple years. The use of a 3-year window was deemed suitable because 3- and 5-year sampling window lengths demonstrated comparable accuracy among estimation methods (as described in appendix 5). With the exception of the North Racoon River near Sac City, Iowa, site (USGS station 05482300, hereafter referred to as "SAC"; 75 percent within ±20 percent of observed loads), continuous monitoring sites had relatively stable streamflow conditions (table 1), and thus nitrate plus nitrite loads at the Mississippi River at Baton Rouge, Louisiana (USGS station 07374000); Illinois River at Valley City, Illinois (USGS station 05586100); POTO; and Connecticut River at Middle Haddam, Connecticut (USGS station 01193050), sites were among the most accurate estimates of the sites considered (92–98 percent within ±20 percent of observed loads).

As with other constituents, the fewest total phosphorus estimates within ±20 percent of observed loads were generally observed at sites with increased variability in daily loading conditions (ROCK, 39 percent; HONE, 50 percent; and SAND, 63 percent). At the two sites with the most variable daily loads (ROCK and HONE), WRTDS_K (63 percent and 78 percent within ±20 percent of observed loads, respectively), WRTDS (53 percent and 75 percent, respectively), and RATIO_F5 (56 percent and 69 percent, respectively) produced the most accurate estimates. Fewer suspended-sediment load estimates were within ±20 percent of observed loads at the RAPP (23 percent within ±20 percent of observed loads), DELA (35 percent), and POTO (36 percent) sites



Comparisons are completed using biweekly and high-flow sampling strategies. ±, plus or minus; MUSK, Muskingum River at McConnelsville, Ohio; GREA, Great Miami River at Miamisburg, Ohio; RAIS, River Raisin at Monroe, Michigan; MAUM, Maumee River at Waterville, Ohio; HONE, Honey Creek at Melmore, Ohio; ROCK, Rock Creek at Tiffin, Ohio; SAND, Sandusky River near Fremont, Ohio; INTERP, interpolation of sampled values; RATIO_T, Beale's ratio estimation with time-based stratification; RATIO_F1, Beale's ratio estimation with flow-based stratification from the target year only; RATIO_F5, Beale's ratio estimation with flow-based stratification on the most recent 5 years of data; L1, LOADEST stock 1-parameter model using data from the target year only; L5, LOADEST stock 5-parameter model; L7, LOADEST stock 7-parameter model; LAICO, LOADEST stock "best selection" model; AIC, LOADEST minimum Akaike information criteria method with cubic flow and flow anomaly variables; PVAL, LOADEST minimum probability-value model with cubic flow and flow anomaly variables; AIC_COMP, LOADEST minimum Akaike information criteria method with local adjustments for residual departures using the composite method; WRTDS, Weighted Regressions on Time, Discharge, and Season; WRTDS_K, Weighted Regressions on Time, Discharge, and Season method with Kalman filtering

Figure 6. Percentage of load estimates within plus or minus 20 percent of observed loads among estimation methods and water-quality constituents.

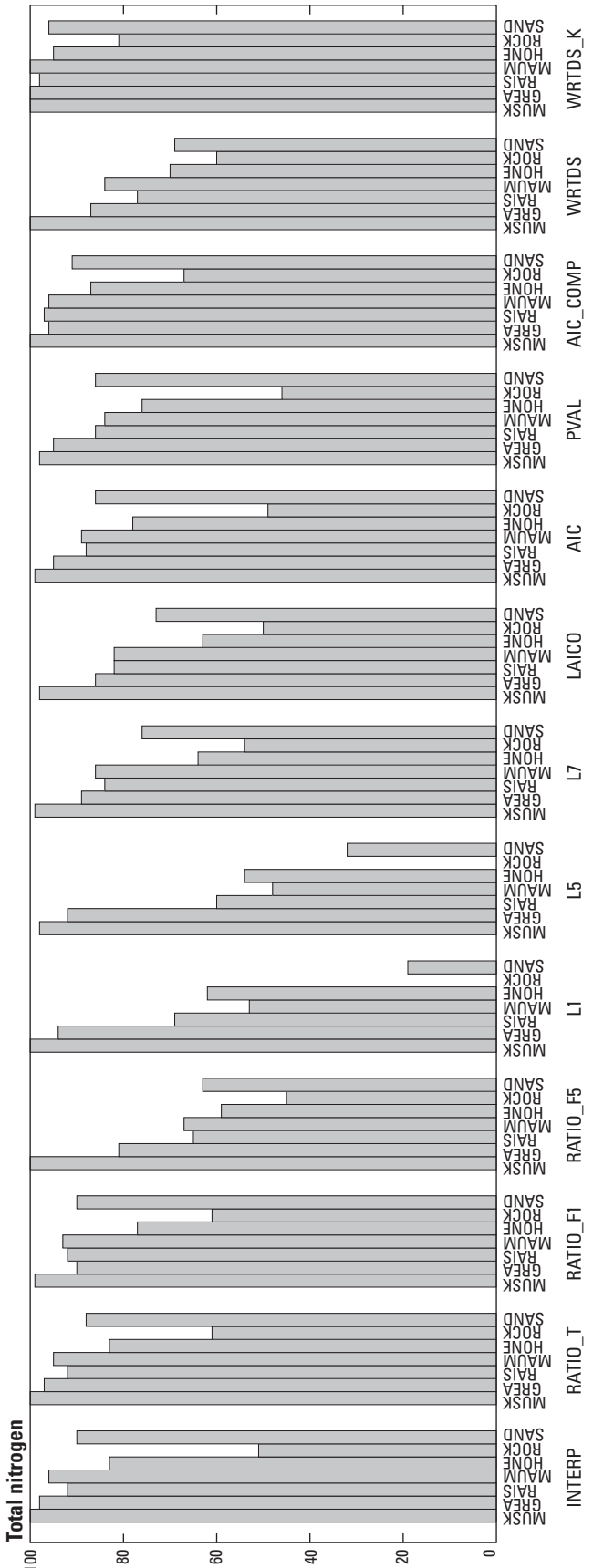
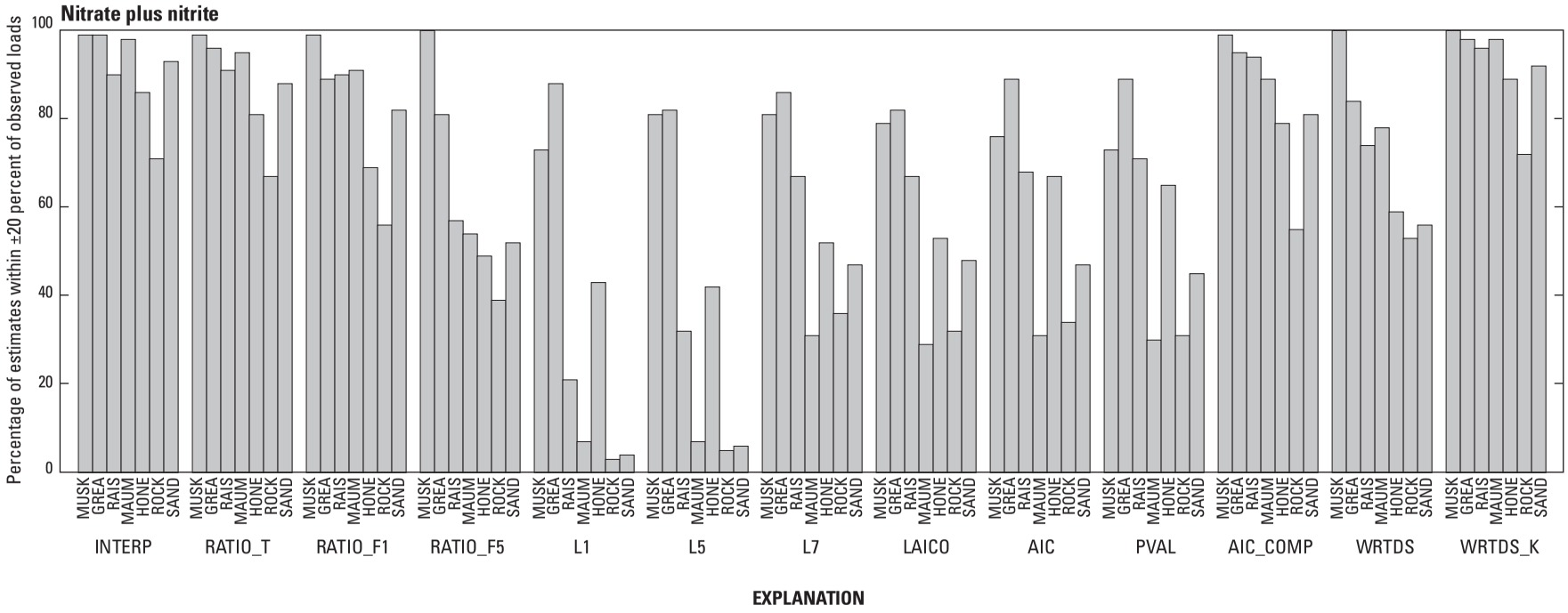


Figure 6. Percentage of load estimates within plus or minus 20 percent of observed loads among estimation methods and water-quality constituents.—Continued



EXPLANATION

Comparisons are completed using biweekly and high-flow sampling strategies. ±, plus or minus; MUSK, Muskingum River at McConnelsville, Ohio; GREA, Great Miami River at Miamisburg, Ohio; RAIS, River Raisin at Monroe, Michigan; MAUM, Maumee River at Waterville, Ohio; HONE, Honey Creek at Melmore, Ohio; ROCK, Rock Creek at Tiffin, Ohio; SAND, Sandusky River near Fremont, Ohio; INTERP, interpolation of sampled values; RATIO_T, Beale's ratio estimation with time-based stratification; RATIO_F1, Beale's ratio estimation with flow-based stratification from the target year only; RATIO_F5, Beale's ratio estimation with flow-based stratification on the most recent 5 years of data; L1, LOADEST stock 1-parameter model using data from the target year only; L5, LOADEST stock 5-parameter model; L7, LOADEST stock 7-parameter model; LAICO, LOADEST stock "best selection" model; AIC, LOADEST minimum Akaike information criteria method with cubic flow and flow anomaly variables; PVAL, LOADEST minimum probability-value model with cubic flow and flow anomaly variables; AIC_COMP, LOADEST minimum Akaike information criteria method with local adjustments for residual departures using the composite method; WRTDS, Weighted Regressions on Time, Discharge, and Season; WRTDS_K, Weighted Regressions on Time, Discharge, and Season method with Kalman filtering

Figure 6. Percentage of load estimates within plus or minus 20 percent of observed loads among estimation methods and water-quality constituents.—Continued

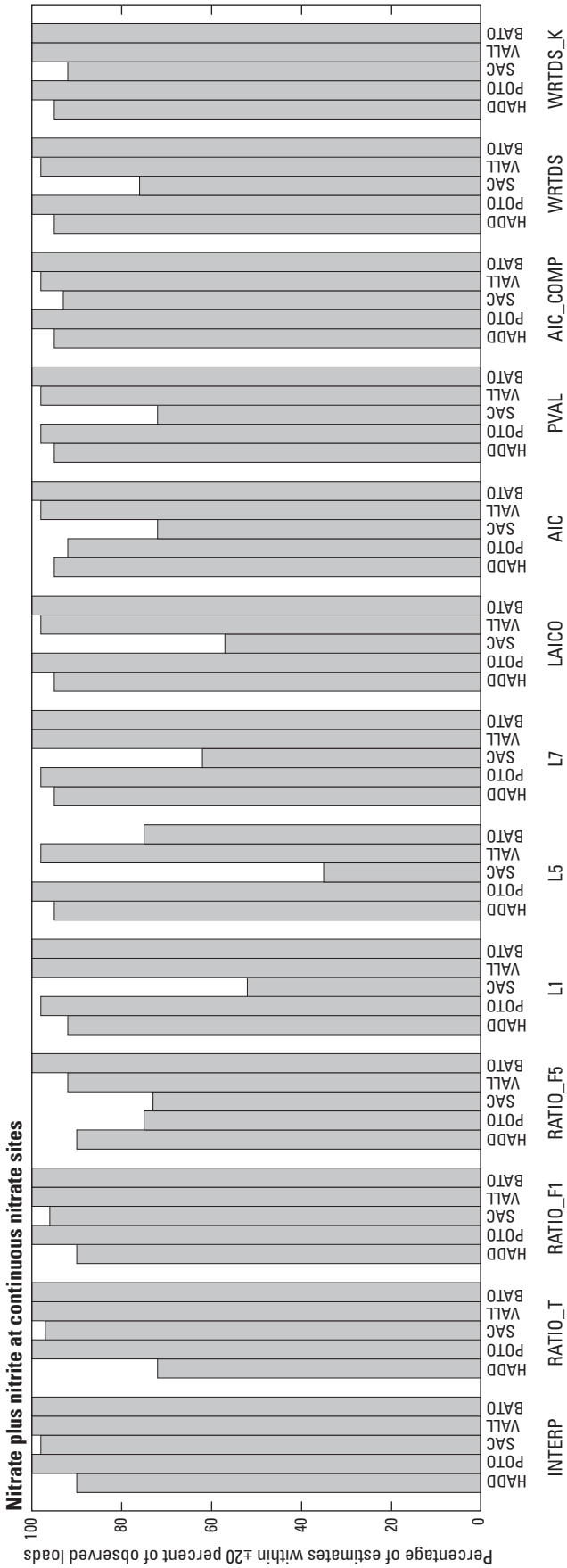


Figure 6. Percentage of load estimates within plus or minus 20 percent of observed loads among estimation methods and water-quality constituents.—Continued

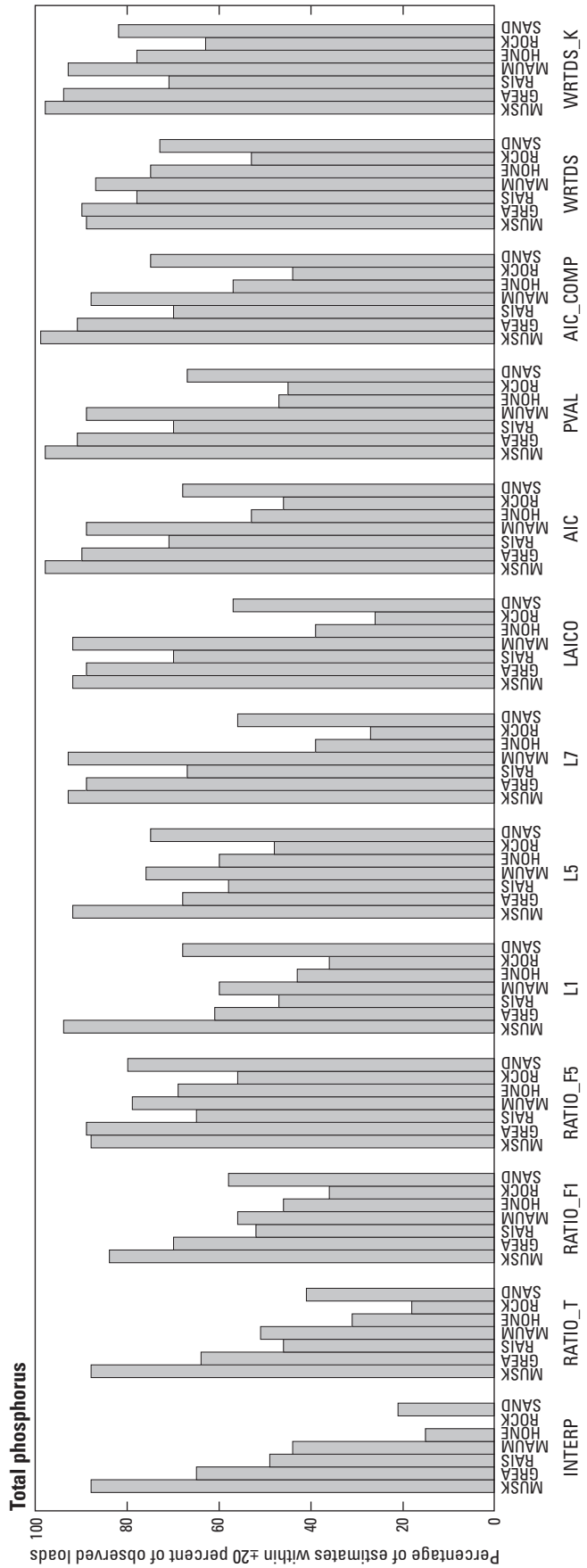
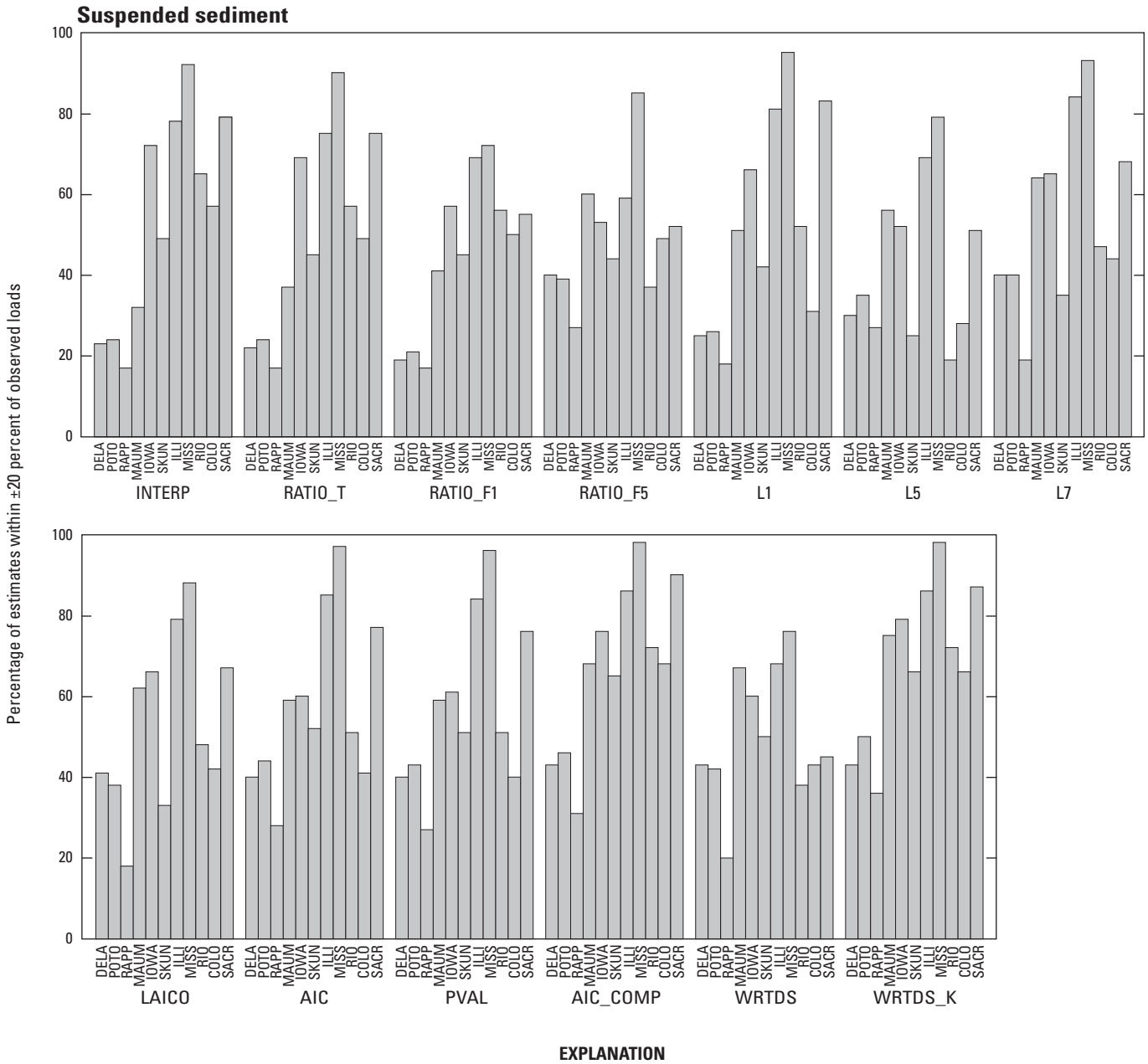


Figure 6. Percentage of load estimates within plus or minus 20 percent of observed loads among estimation methods and water-quality constituents.—Continued



EXPLANATION

Comparisons are completed using biweekly and high-flow sampling strategies. ±, plus or minus; DELA, Delaware River at Trenton, New Jersey; POTO, Potomac River near Washington, D.C., Little Falls Pump Station; RAPP, Rappahannock River at Remington, Virginia; MAUM, Maumee River at Waterville, Ohio; IOWA, Iowa River at Wapello, Iowa; SKUN, Skunk River at Augusta, Iowa; ILLI, Illinois River at Valley City, Illinois; MISS, Mississippi River at Thebes, Ill.; RIO, Rio Grande at Otowi Bridge, New Mexico; COLO, Colorado River near Cisco, Utah; SACR, Sacramento River at Freeport, California; INTERP, interpolation of sampled values; RATIO_T, Beale's ratio estimation with time-based stratification; RATIO_F1, Beale's ratio estimation with flow-based stratification from the target year only; RATIO_F5, Beale's ratio estimation with flow-based stratification on the most recent 5 years of data; L1, LOADEST stock 1-parameter model using data from the target year only; L5, LOADEST stock 5-parameter model ; L7, LOADEST stock 7-parameter model; LAICO, LOADEST stock "best selection" model; AIC, LOADEST minimum Akaike information criteria method with cubic flow and flow anomaly variables; PVAL, LOADEST minimum probability-value model with cubic flow and flow anomaly variables; AIC_COMP, LOADEST minimum Akaike information criteria method with local adjustments for residual departures using the composite method; WRTDS, Weighted Regressions on Time, Discharge, and Season; WRTDS_K, Weighted Regressions on Time, Discharge, and Season method with Kalman filtering

Figure 6. Percentage of load estimates within plus or minus 20 percent of observed loads among estimation methods and water-quality constituents.—Continued

than any other combination of sites and constituents studied herein (fig. 5). WRTDS_K and AIC_COMP (36 percent and 31 percent, respectively) produced the most estimates within the ±20-percent threshold at RAPP, improving upon methods that do not adjust daily estimates based on sampled values (WRTDS, 20 percent; AIC, 28 percent). In contrast, a variety of methods, including WRTDS_K, WRTDS, AIC_COMP, RATIO_F5, L7, LAICO, AIC, and PVAL (40–43 percent within ±20 percent of observed loads) produced estimates of similar accuracy for the DELA site (fig. 6). POTO estimates were most often within ±20 percent of observed loads using the WRTDS_K (50 percent) and AIC_COMP (46 percent) methods, although the AIC, WRTDS, and PVAL methods demonstrated somewhat similar accuracy (42–43 percent within ±20 percent of observed loads). In contrast to other sites, suspended-sediment loads were estimated relatively accurately by multiple methods at the Mississippi River at Thebes, Ill. (USGS station 07022000, hereafter referred to as "MISS"; 89 percent of all estimates within criteria), and Illinois River at Valley City, Ill. (USGS station 05586100, hereafter referred to as "ILLI"; 77 percent of all estimates within criteria), sites, likely because these sites had among the least variable daily loading conditions (fig. 5). LOADEST-based methods (with the exception of L5) and the WRTDS_K method produced the most estimates within the ±20-percent threshold at these sites (MISS, 93–98 percent; ILLI, 81–86 percent).

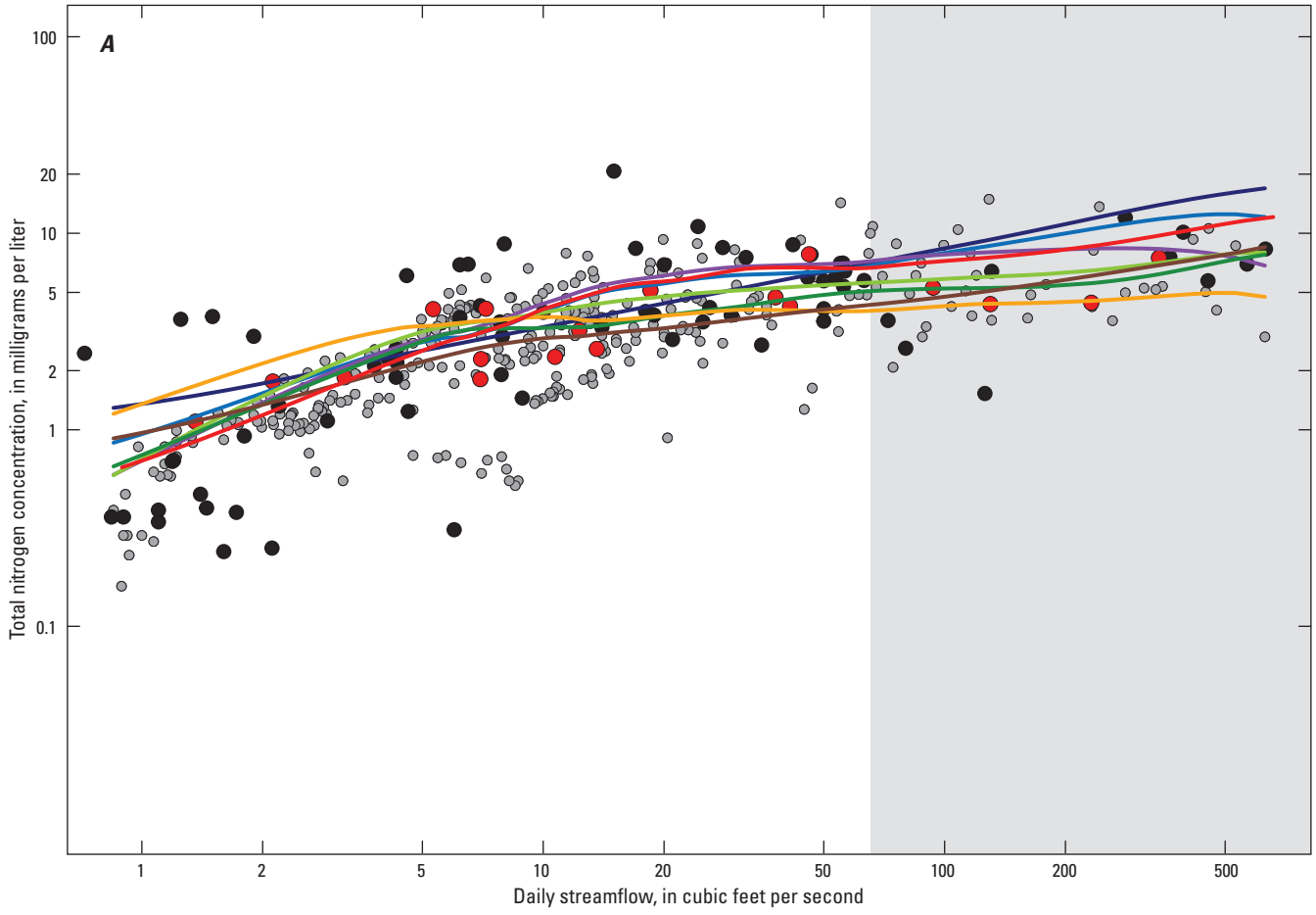
Results presented in figures 5 and 6 indicate that site-specific transport processes often dictate the ability to compute accurate water-quality loads. Sites with more variable loading conditions are more difficult to estimate, in part because few samples are typically collected during the relatively few days that transport most of the annual water-quality load. Site-specific transport processes also may not be adequately mimicked by methods that use static linear, quadratic, or cubic relations with streamflow, season, or time. Although method performance differed among sites and constituents, methods that adjusted daily estimates based on departures from observed values (WRTDS_K and, to a lesser degree, AIC_COMP) generally produced the most accurate estimates, including at sampling sites with more variable loading conditions. However, the variability observed in method performance indicates that even though some methods produce inaccurate results generally, they may work well to model transport processes at specific sampling sites. To better characterize underlying causes of bias in water-quality load estimation, the following section illustrates examples of how estimation methods represent daily water-quality concentrations and loads across streamflow conditions.

Examples of Method Performance

Estimation errors can be generally attributed to biased sampling, model misspecification, or errors in the retransformation bias correction caused by heteroscedastic error distributions. Examples of how methods estimate daily total nitrogen, nitrate plus nitrite, total phosphorus, and suspended-sediment concentrations and loads relative to streamflow conditions at two sites (ROCK and POTO) are presented in figures 7–10. These figures are formatted to illustrate (1) how selected sampling strategies compare to observed daily concentrations and loads and (2) how assumptions inherent in load-estimation methods relate to observed values across streamflow conditions. Chloride is not considered in this section because most methods were able to produce relatively accurate load estimates. Ratio estimators also are not illustrated because they do not produce daily estimates, and LAICO and PVAL estimates are not illustrated because estimates from these methods typically differ little from other regression-based methods (L5 or L7 in the case of LAICO, and AIC in the case of PVAL).

Some explanation is necessary to clarify the multiple types of information depicted in figures 7–10. Examples illustrate general patterns in estimation method performance relative to daily streamflow conditions for a specific water-quality constituent, sampling site, sampling strategy, water year, and replicate. The sites, constituents, water years, and replicates represented in these figures are indicative of broad patterns in the estimation method performance presented previously. Observed, sampled, and estimated water-quality concentrations across streamflow conditions in logarithmic space are compared in figures 7*A*, 8*A*, 9*A*, and 10*A*; daily water-quality loads in relation to streamflow conditions for the same site, constituent, water year, and replicate in arithmetic space are shown in figures 7*B*, 8*B*, 9*B*, and 10*B*. Because daily estimates from multiple methods would be impossible to discern, estimation method performance is characterized by a loess fit of daily estimates relative to streamflow conditions. This approach illustrates the general response of methods across streamflow conditions and relative to sampled and observed concentrations and loads. Method performance is summarized generally and specifically for high streamflows that transport 80 percent of the annual load (figs. 7–10). The same site (ROCK) and water year (2004) were illustrated for total nitrogen, nitrate plus nitrite, and total phosphorus examples to illustrate differences in constituent transport and method performance during identical streamflow conditions. Because daily suspended-sediment values were not collected at ROCK, POTO was selected to illustrate an example of method performance for computing suspended-sediment loads. Method performance is summarized for high streamflows specifically because misspecification of concentrations during high-flows can result in biased load estimates for the entire year, although methods may adequately represent mean concentrations throughout the year.

Observed, sampled, and estimated daily total nitrogen concentrations and loads computed using HIFLOW sampling at ROCK in 2004 are presented in figure 7. Slightly more than 80 percent of the loads in this example were transported during 41 days (out of a possible 334) in which streamflows were greater than 65 cubic feet per second (ft^3/s). The ability



EXPLANATION

[Loess, locally estimated scatterplot smoothing; INTERP, interpolation of sampled values; L1, LOADEST stock 1-parameter model using data from the target year only; L5, LOADEST stock 5-parameter model ; L7, LOADEST stock 7-parameter model; AIC, LOADEST minimum Akaike information criteria method with cubic flow and flow anomaly variables; AIC_COMP, LOADEST minimum Akaike information criteria method with local adjustments for residual departures using the composite method; WRTDS, Weighted Regressions on Time, Discharge, and Season; WRTDS_K, Weighted Regressions on Time, Discharge, and Season method with Kalman filtering; mg/L, milligram per liter]

Summary of concentrations across all streamflow conditions

Observed mean: 3.30 mg/L
Sampled mean target year: 3.81 mg/L
Sampled mean 5-year window: 4.21 mg/L
INTERP: 3.44 mg/L
L1: 4.06 mg/L
L5: 4.30 mg/L
L7: 4.30 mg/L
AIC: 4.21 mg/L
AIC_COMP: 2.89 mg/L
WRTDS: 3.79 mg/L
WRTDS_K: 3.27 mg/L

Indicates streamflow conditions that transport 80 percent of the annual load

Loess of method estimates

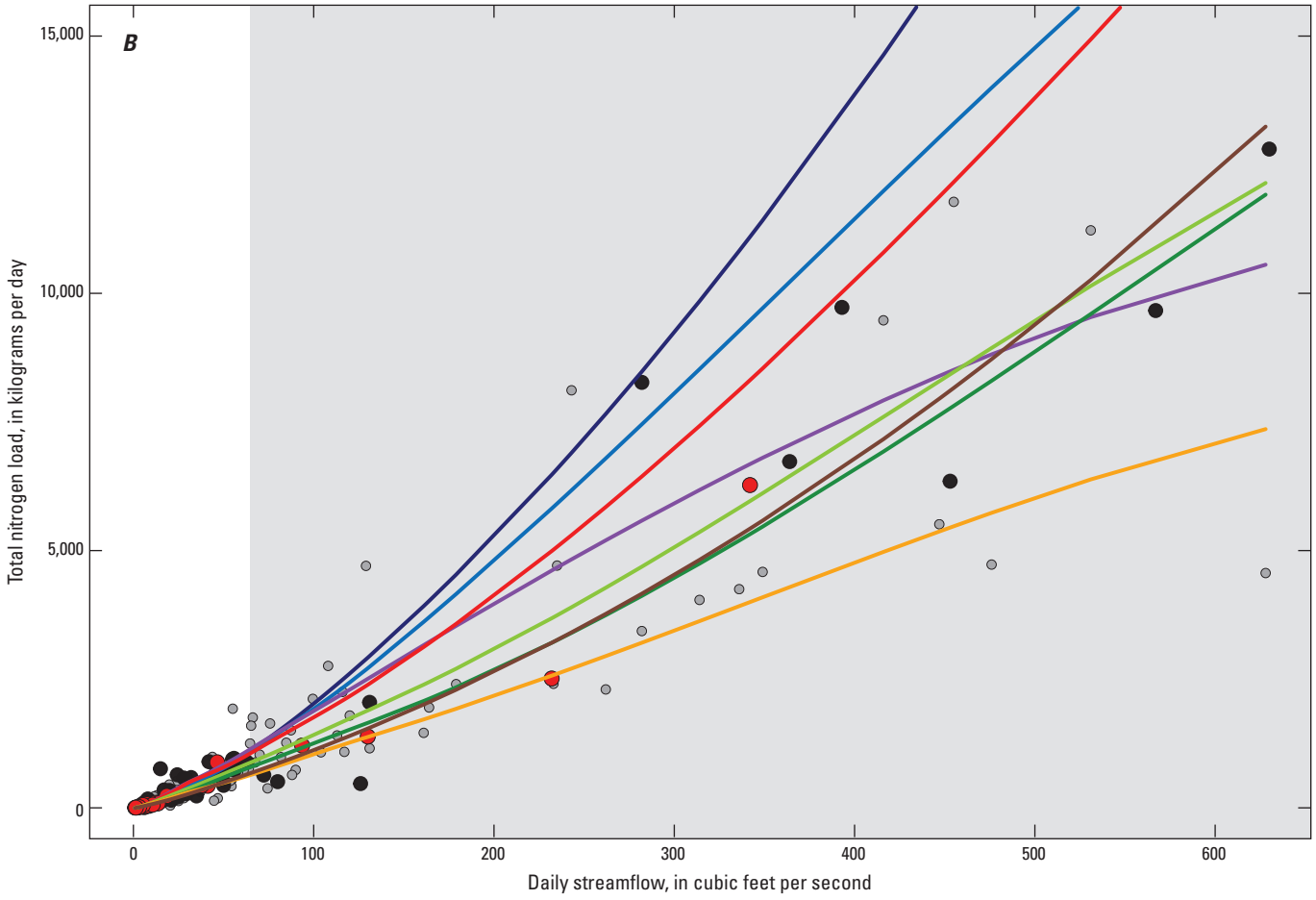
INTERP
L1
L5
L7
AIC
AIC_COMP
WRTDS
WRTDS_K

Observed value
Sampled value in target year
Sampled value in 5-year window

Summary of concentrations during streamflows that transported 80 percent of the annual load

Observed mean: 6.32 mg/L
Sampled mean target year: 5.40 mg/L
Sampled mean 5-year window: 6.17 mg/L
INTERP: 4.63 mg/L
L1: 10.4 mg/L
L5: 9.31 mg/L
L7: 7.78 mg/L
AIC: 8.49 mg/L
AIC_COMP: 5.51 mg/L
WRTDS: 6.40 mg/L
WRTDS_K: 5.70 mg/L

Figure 7. Observed, sampled, and estimated total nitrogen collected using the high-flow sampling strategy at the Rock Creek at Tiffin, Ohio, site (U.S. Geological Survey station 04197170) in 2004. *A*, Total nitrogen concentrations relative to streamflow conditions in logarithmic space; and *B*, total nitrogen loads relative to streamflow conditions in arithmetic space.



EXPLANATION

[Loess, locally estimated scatterplot smoothing; INTERP, interpolation of sampled values; L1, LOADEST stock 1-parameter model using data from the target year only; L5, LOADEST stock 5-parameter model ; L7, LOADEST stock 7-parameter model; AIC, LOADEST minimum Akaike information criteria method with cubic flow and flow anomaly variables; AIC_COMP, LOADEST minimum Akaike information criteria method with local adjustments for residual departures using the composite method; WRTDS, Weighted Regressions on Time, Discharge, and Season; WRTDS_K, Weighted Regressions on Time, Discharge, and Season method with Kalman filtering; kg/day, kilogram per day]

Summary of loads and estimation method performance across all streamflow conditions

Observed mean: 488 kg/d
Sampled mean target year: 769 kg/d
Sampled mean 5-year window: 960 kg/d
INTERP: –20 percent error
L1: 80 percent error
L5: 59 percent error
L7: 26 percent error
AIC: 52 percent error
AIC_COMP: 47 percent error
WRTDS: 11 percent error
WRTDS_K: 1 percent error

Indicates streamflow conditions that transport 80 percent of the annual load

Loess of method estimates

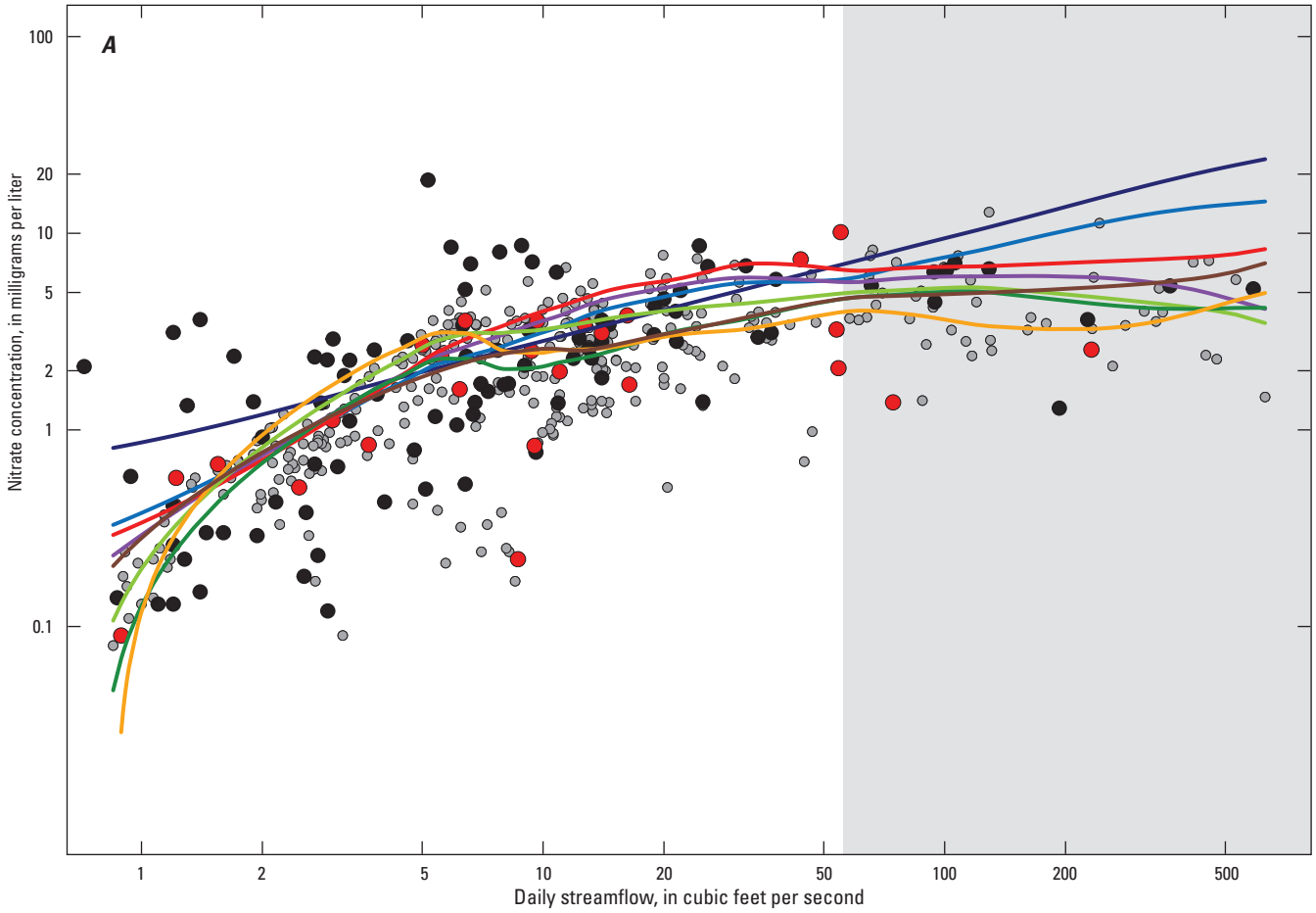
- INTERP
- L1
- L5
- L7
- AIC
- AIC_COMP
- WRTDS
- WRTDS_K

Observed value
Sampled value in target year
Sampled value in 5-year window

Summary of loads and estimation method performance during streamflows that transported 80 percent of the annual load

Observed mean: 3,149 kg/day
Sampled mean target year: 2,847 kg/day
Sampled mean 5-year window: 4,901 kg/day
INTERP: –23 percent error
L1: 98 percent error
L5: 68 percent error
L7: 24 percent error
AIC: 52 percent error
AIC_COMP: 0.2 percent error
WRTDS: 11 percent error
WRTDS_K: 0.3 percent error

Figure 7. Observed, sampled, and estimated total nitrogen collected using the high-flow sampling strategy at the Rock Creek at Tiffin, Ohio, site (U.S. Geological Survey station 04197170) in 2004. *A*, Total nitrogen concentrations relative to streamflow conditions in logarithmic space; and *B*, total nitrogen loads relative to streamflow conditions in arithmetic space.—Continued



EXPLANATION

[Loess, locally estimated scatterplot smoothing; INTERP, interpolation of sampled values; L1, LOADEST stock 1-parameter model using data from the target year only; L5, LOADEST stock 5-parameter model ; L7, LOADEST stock 7-parameter model; AIC, LOADEST minimum Akaike information criteria method with cubic flow and flow anomaly variables; AIC_COMP, LOADEST minimum Akaike information criteria method with local adjustments for residual departures using the composite method; WRTDS, Weighted Regressions on Time, Discharge, and Season; WRTDS_K, Weighted Regressions on Time, Discharge, and Season method with Kalman filtering; mg/L, milligram per day]

Summary of concentrations across all streamflow conditions

Observed mean: 2.50 mg/L
Sampled mean target year: 2.48 mg/L
Sampled mean 5-year window: 2.80 mg/L
INTERP: 2.53 mg/L
L1: 3.97 mg/L
L5: 3.59 mg/L
L7: 3.36 mg/L
AIC: 3.79 mg/L
AIC_COMP: 2.52 mg/L
WRTDS: 3.02 mg/L
WRTDS_K: 2.40 mg/L

Indicates streamflow conditions that transport 80 percent of the annual load

Loess of method estimates

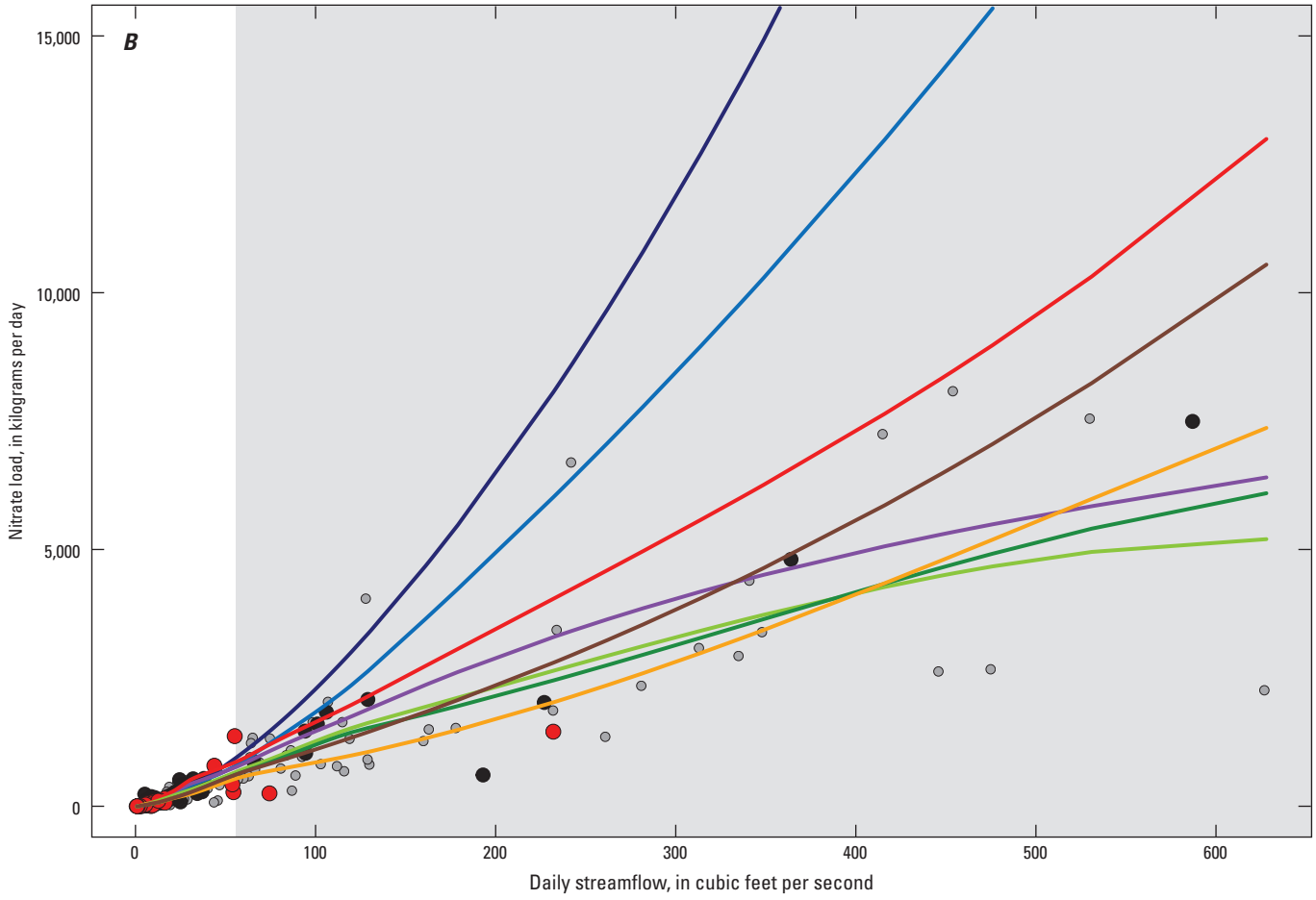
- INTERP
- L1
- L5
- L7
- AIC
- AIC_COMP
- WRTDS
- WRTDS_K

Observed value
Sampled value in target year
Sampled value in 5-year window

Summary of concentrations during streamflows that transported 80 percent of the annual load

Observed mean: 4.58 mg/L
Sampled mean target year: 1.97 mg/L
Sampled mean 5-year window: 4.66 mg/L
INTERP: 3.85 mg/L
L1: 12.30 mg/L
L5: 9.11 mg/L
L7: 5.67 mg/L
AIC: 6.98 mg/L
AIC_COMP: 4.81 mg/L
WRTDS: 4.92 mg/L
WRTDS_K: 4.73 mg/L

Figure 8. Observed, sampled, and estimated nitrate plus nitrite collected using the biweekly sampling strategy at the Rock Creek at Tiffin, Ohio, site (U.S. Geological Survey station 04197170) in 2004. *A*, Nitrate plus nitrite concentrations relative to streamflow conditions in logarithmic space; and *B*, nitrate plus nitrite loads relative to streamflow conditions in arithmetic space.



EXPLANATION

[Loess, locally estimated scatterplot smoothing; INTERP, interpolation of sampled values; L1, LOADEST stock 1-parameter model using data from the target year only; L5, LOADEST stock 5-parameter model ; L7, LOADEST stock 7-parameter model; AIC, LOADEST minimum Akaike information criteria method with cubic flow and flow anomaly variables; AIC_COMP, LOADEST minimum Akaike information criteria method with local adjustments for residual departures using the composite method; WRTDS, Weighted Regressions on Time, Discharge, and Season; WRTDS_K, Weighted Regressions on Time, Discharge, and Season method with Kalman filtering; kg/day, kilogram per day]

Summary of loads and estimation method performance across all streamflow conditions

Observed mean: 349 kg/d
Sampled mean target year: 215 kg/d
Sampled mean 5-year window: 298 kg/d
INTERP: –6.9 percent error
L1: 212 percent error
L5: 123 percent error
L7: 24 percent error
AIC: 63 percent error
AIC_COMP: 15 percent error
WRTDS: 7.4 percent error
WRTDS_K: 1.7 percent error

Indicates streamflow conditions that transport 80 percent of the annual load

Loess of method estimates

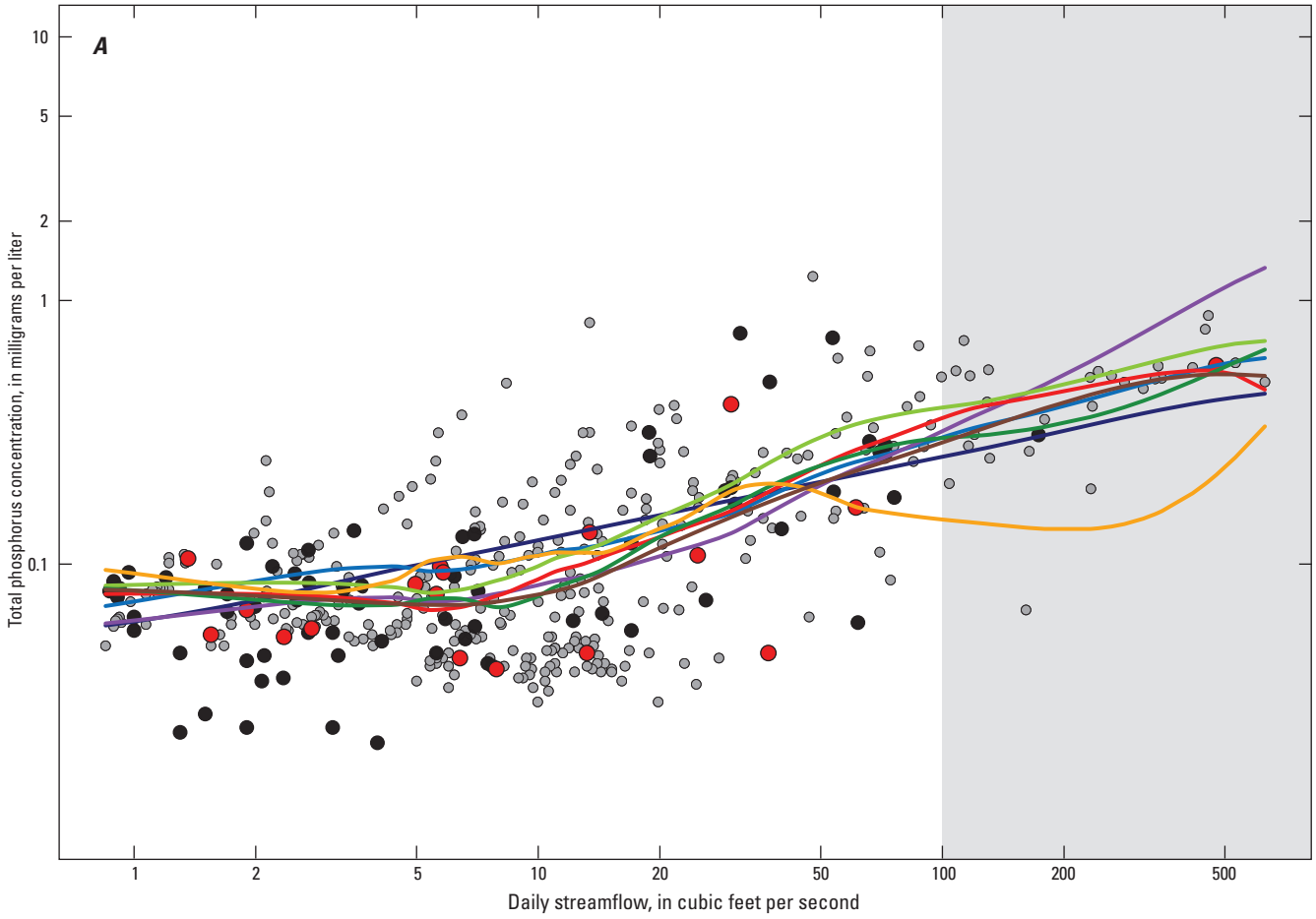
- INTERP
- L1
- L5
- L7
- AIC
- AIC_COMP
- WRTDS
- WRTDS_K

Observed value
Sampled value in target year
Sampled value in 5-year window

Summary of loads and estimation method performance during streamflows that transported 80 percent of the annual load

Observed mean: 2,073 kg/day
Sampled mean target year: 2,162 kg/day
Sampled mean 5-year window: 852 kg/day
INTERP: –7.5 percent error
L1: 258 percent error
L5: 147 percent error
L7: 20 percent error
AIC: 64 percent error
AIC_COMP: 19 percent error
WRTDS: 4.1 percent error
WRTDS_K: 3.3 percent error

Figure 8. Observed, sampled, and estimated nitrate plus nitrite collected using the biweekly sampling strategy at the Rock Creek at Tiffin, Ohio, site (U.S. Geological Survey station 04197170) in 2004. *A*, Nitrate plus nitrite concentrations relative to streamflow conditions in logarithmic space; and *B*, nitrate plus nitrite loads relative to streamflow conditions in arithmetic space.—Continued



EXPLANATION

[Loess, locally estimated scatterplot smoothing; INTERP, interpolation of sampled values; L1, LOADEST stock 1-parameter model using data from the target year only; L5, LOADEST stock 5-parameter model ; L7, LOADEST stock 7-parameter model; AIC, LOADEST minimum Akaike information criteria method with cubic flow and flow anomaly variables; AIC_COMP, LOADEST minimum Akaike information criteria method with local adjustments for residual departures using the composite method; WRTDS, Weighted Regressions on Time, Discharge, and Season; WRTDS_K, Weighted Regressions on Time, Discharge, and Season method with Kalman filtering; mg/L, milligram per liter]

Summary of concentrations across all streamflow conditions

Observed mean: 0.15 mg/L
Sampled mean target year: 0.12 mg/L
Sampled mean 5-year window: 0.12 mg/L
INTERP: 0.12 mg/L
L1: 0.14 mg/L
L5: 0.14 mg/L
L7: 0.14 mg/L
AIC: 0.14 mg/L
AIC_COMP: 0.12 mg/L
WRTDS: 0.16 mg/L
WRTDS_K: 0.13 mg/L

Indicates streamflow conditions that transport 80 percent of the annual load

Loess of method estimates

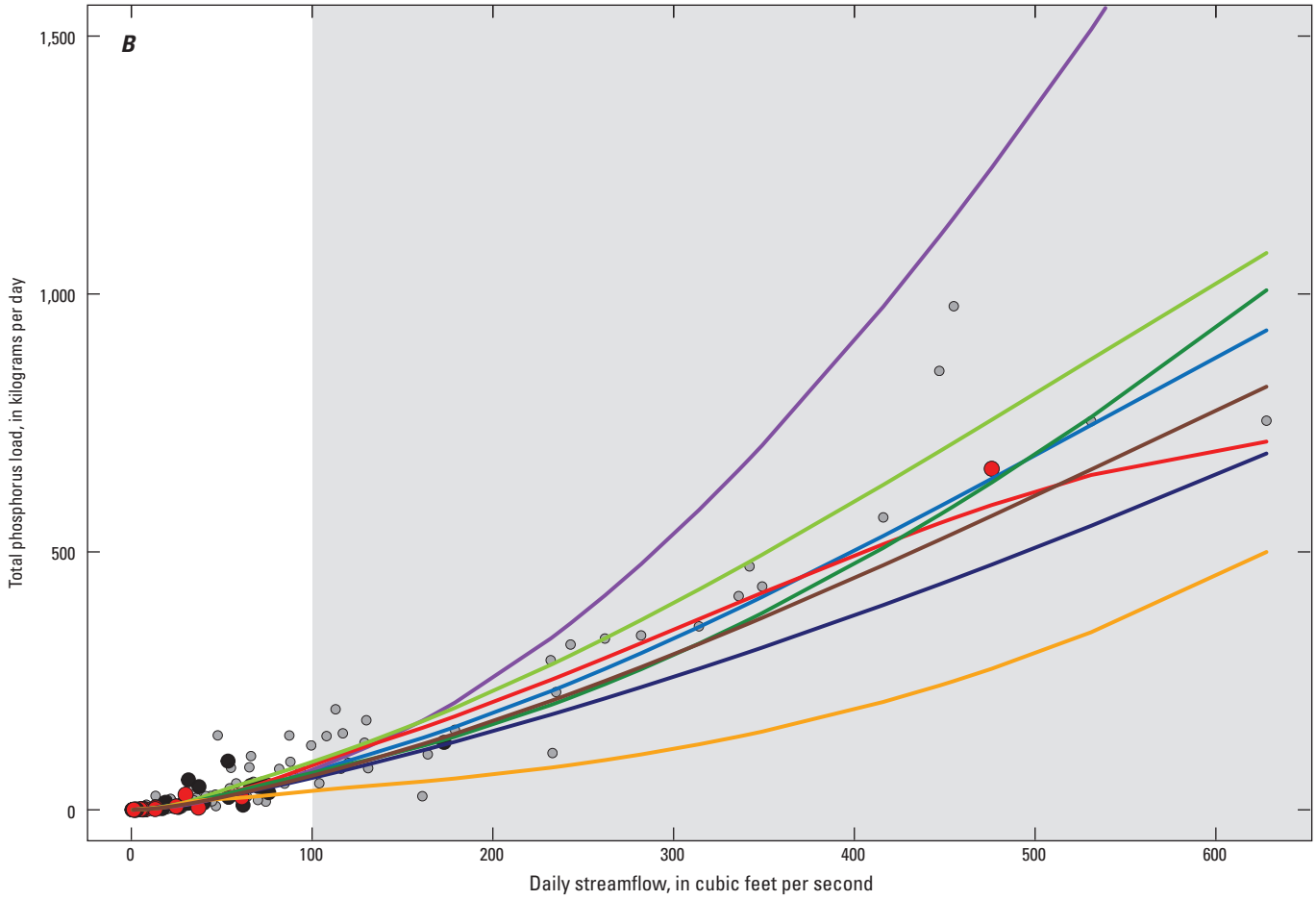
- INTERP
- L1
- L5
- L7
- AIC
- AIC_COMP
- WRTDS
- WRTDS_K

Observed value
Sampled value in target year
Sampled value in 5-year window

Summary of concentrations during streamflows that transported 80 percent of the annual load

Observed mean: 0.46 mg/L
Sampled mean target year: 0.57 mg/L
Sampled mean 5-year window: 0.44 mg/L
INTERP: 0.16 mg/L
L1: 0.33 mg/L
L5: 0.43 mg/L
L7: 0.66 mg/L
AIC: 0.46 mg/L
AIC_COMP: 0.40 mg/L
WRTDS: 0.51 mg/L
WRTDS_K: 0.39 mg/L

Figure 9. Observed, sampled, and estimated total phosphorus collected using the high-flow sampling strategy at the Rock Creek at Tiffin, Ohio, site (U.S. Geological Survey station 04197170) in 2004. *A*, Total phosphorus concentrations relative to streamflow conditions in logarithmic space; and *B*, total phosphorus loads relative to streamflow conditions in arithmetic space.



EXPLANATION

[Loess, locally estimated scatterplot smoothing; INTERP, interpolation of sampled values; L1, LOADEST stock 1-parameter model using data from the target year only; L5, LOADEST stock 5-parameter model ; L7, LOADEST stock 7-parameter model; AIC, LOADEST minimum Akaike information criteria method with cubic flow and flow anomaly variables; AIC_COMP, LOADEST minimum Akaike information criteria method with local adjustments for residual departures using the composite method; WRTDS, Weighted Regressions on Time, Discharge, and Season; WRTDS_K, Weighted Regressions on Time, Discharge, and Season method with Kalman filtering; kg/day, kilogram per day]

Summary of loads and estimation method performance across all streamflow conditions

Observed mean: 35 kg/d
Sampled mean target year: 41 kg/d
Sampled mean 5-year window: 16 kg/d
INTERP: –58 percent error
L1: –27 percent error
L5: –8.9 percent error
L7: 42 percent error
AIC: –6.8 percent error
AIC_COMP: –17 percent error
WRTDS: 8.4 percent error
WRTDS_K: –13 percent error

Indicates streamflow conditions that transport 80 percent of the annual load

Loess of method estimates

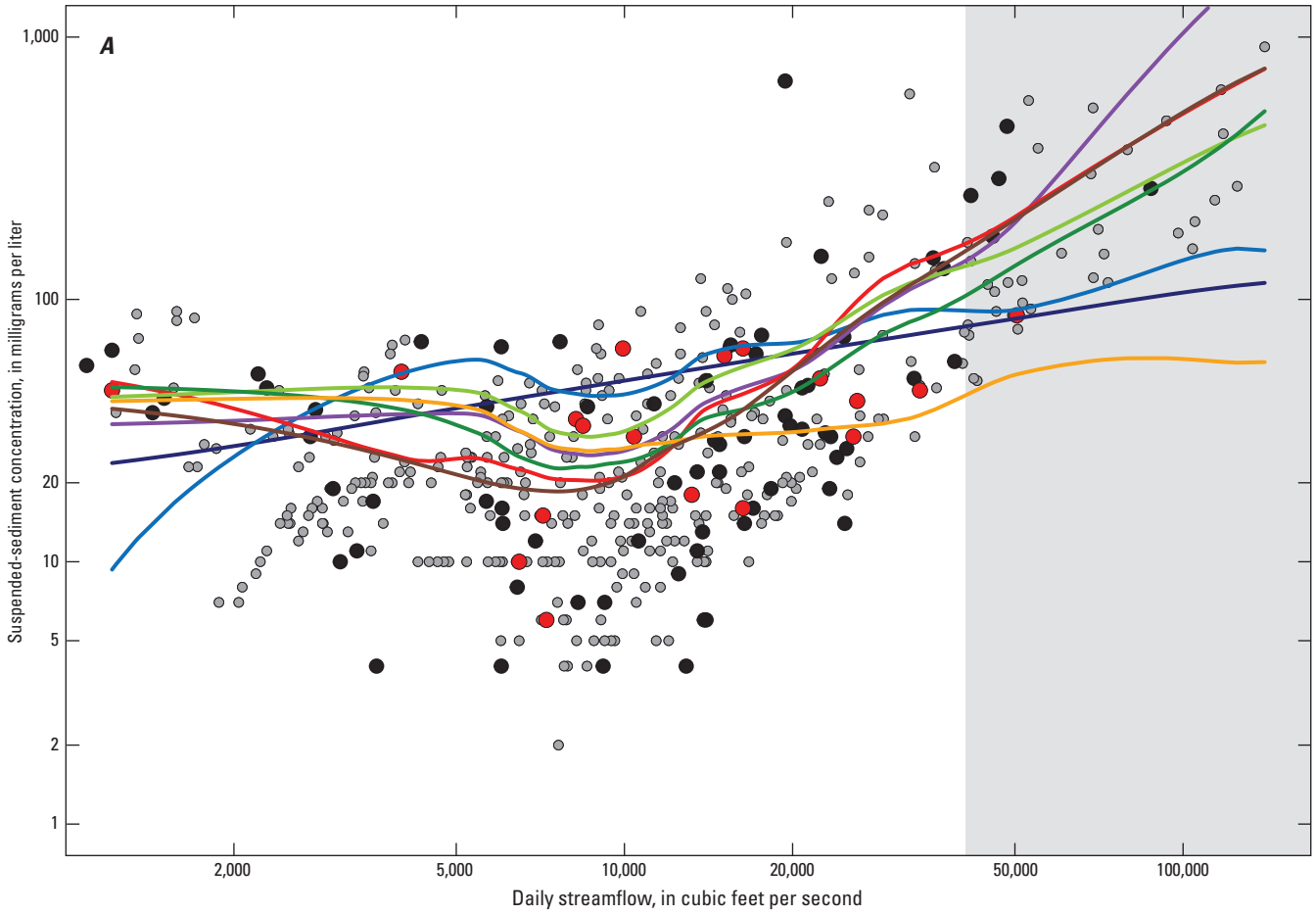
- INTERP
- L1
- L5
- L7
- AIC
- AIC_COMP
- WRTDS
- WRTDS_K

- **Observed value**
- **Sampled value in target year**
- **Sampled value in 5-year window**

Summary of loads and estimation method performance during streamflows that transported 80 percent of the annual load

Observed mean: 330 kg/day
Sampled mean target year: 661 kg/day
Sampled mean 5-year window: 396 kg/day
INTERP: –64 percent error
L1: –29 percent error
L5: –6.3 percent error
L7: 59 percent error
AIC: –4.5 percent error
AIC_COMP: –1 percent error
WRTDS: 10 percent error
WRTDS_K: –12 percent error

Figure 9. Observed, sampled, and estimated total phosphorus collected using the high-flow sampling strategy at the Rock Creek at Tiffin, Ohio, site (U.S. Geological Survey station 04197170) in 2004. *A*, Total phosphorus concentrations relative to streamflow conditions in logarithmic space; and *B*, total phosphorus loads relative to streamflow conditions in arithmetic space.—Continued



EXPLANATION

[Loess, locally estimated scatterplot smoothing; INTERP, interpolation of sampled values; L1, LOADEST stock 1-parameter model using data from the target year only; L5, LOADEST stock 5-parameter model ; L7, LOADEST stock 7-parameter model; AIC, LOADEST minimum Akaike information criteria method with cubic flow and flow anomaly variables; AIC_COMP, LOADEST minimum Akaike information criteria method with local adjustments for residual departures using the composite method; WRTDS, Weighted Regressions on Time, Discharge, and Season; WRTDS_K, Weighted Regressions on Time, Discharge, and Season method with Kalman filtering; mg/L, milligram per liter]

Summary of concentrations across all streamflow conditions

Observed mean: 53 mg/L
Sampled mean target year: 39 mg/L
Sampled mean 5-year window: 57 mg/L
INTERP: 35 mg/L
L1: 51 mg/L
L5: 58 mg/L
L7: 90 mg/L
AIC: 65 mg/L
AIC_COMP: 66 mg/L
WRTDS: 65 mg/L
WRTDS_K: 52 mg/L

Indicates streamflow conditions that transport 80 percent of the annual load

Loess of method estimates

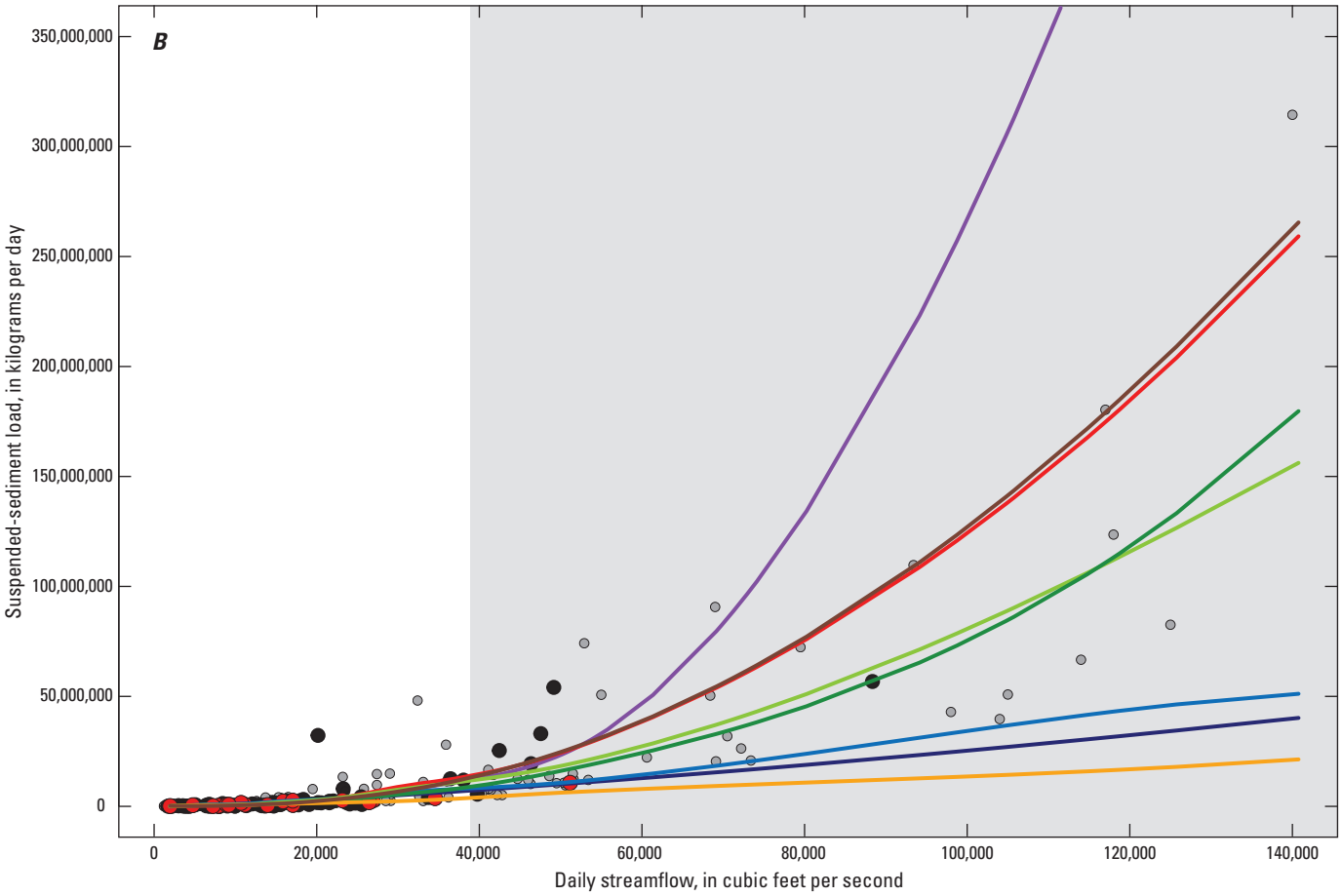
INTERP
L1
L5
L7
AIC
AIC_COMP
WRTDS
WRTDS_K

Observed value
Sampled value in target year
Sampled value in 5-year window

Summary of concentrations during streamflows that transported 80 percent of the annual load

Observed mean: 226 mg/L
Sampled mean target year: 87 mg/L
Sampled mean 5-year window: 254 mg/L
INTERP: 53 mg/L
L1: 92 mg/L
L5: 110 mg/L
L7: 561 mg/L
AIC: 321 mg/L
AIC_COMP: 353 mg/L
WRTDS: 225 mg/L
WRTDS_K: 196 mg/L

Figure 10. Observed, sampled, and estimated suspended sediment collected using the high-flow sampling strategy at the Potomac River near Washington, D.C., Little Falls Pump Station site (U.S. Geological Survey station 01646500) in 1978. *A*, Suspended-sediment concentrations relative to streamflow conditions in logarithmic space; and *B*, suspended-sediment loads relative to streamflow conditions in arithmetic space.



EXPLANATION

[Loess, locally estimated scatterplot smoothing; INTERP, interpolation of sampled values; L1, LOADEST stock 1-parameter model using data from the target year only; L5, LOADEST stock 5-parameter model ; L7, LOADEST stock 7-parameter model; AIC, LOADEST minimum Akaike information criteria method with cubic flow and flow anomaly variables; AIC_COMP, LOADEST minimum Akaike information criteria method with local adjustments for residual departures using the composite method; WRTDS, Weighted Regressions on Time, Discharge, and Season; WRTDS_K, Weighted Regressions on Time, Discharge, and Season method with Kalman filtering; kg/day, kilogram per day]

Summary of loads and estimation method performance across all streamflow conditions

Observed mean: 5.64×10^{6} kg/d
Sampled mean target year: 1.81×10^{6} kg/d
Sampled mean 5-year window: 3.85×10^{6} kg/d
INTERP: –71 percent error
L1: –49 percent error
L5: –38 percent error
L7: 147 percent error
AIC: 38 percent error
AIC_COMP: 45 percent error
WRTDS: 2.3 percent error
WRTDS_K: –13 percent error

Indicates streamflow conditions that transport 80 percent of the annual load

Loess of method estimates

- INTERP
- L1
- L5
- L7
- AIC
- AIC_COMP
- WRTDS
- WRTDS_K

Observed value
Sampled value in target year
Sampled value in 5-year window

Summary of loads and estimation method performance during streamflows that transported 80 percent of the annual load

Observed mean: 4.70×10^{7} kg/day
Sampled mean target year: 1.07×10^{7} kg/day
Sampled mean 5-year window: 3.32×10^{7} kg/day
INTERP: –81 percent error
L1: –65 percent error
L5: –55 percent error
L7: 182 percent error
AIC: 43 percent error
AIC_COMP: 53 percent error
WRTDS: –3.3 percent error
WRTDS_K: –14 percent error

Figure 10. Observed, sampled, and estimated suspended sediment collected using the high-flow sampling strategy at the Potomac River near Washington, D.C., Little Falls Pump Station site (U.S. Geological Survey station 01646500) in 1978. *A*, Suspended-sediment concentrations relative to streamflow conditions in logarithmic space; and *B*, suspended-sediment loads relative to streamflow conditions in arithmetic space.—Continued

of samples to represent observed conditions can vary among the whole record and during specific streamflow conditions, as shown in figure 7*A*. Mean sampled concentrations during 2004 and the previous 4 years were slightly larger than observed values generally but were somewhat smaller than observed values at streamflows above 65 ft^3/s (fig. 7*A*).

Illustrated in figures 7*A* and *B* is the way in which underlying assumptions (or lack of assumptions) of how concentrations and loads change relative to streamflow conditions affect estimation method performance. Mean INTERP estimates of total nitrogen concentration (3.44 milligrams per liter [mg/L]) were similar to observed values (3.30 mg/L) across all streamflow conditions but were biased low during high streamflows (fig. 7*A*). Underrepresentation of observed values during higher streamflows caused INTERP to produce an annual load estimate that was biased low (−20 percent; fig. 7*B*), which corresponds to patterns observed for INTERP estimates of total nitrogen for other sites and years (fig. 3). Because total nitrogen concentrations are nonlinear with respect to streamflow conditions in logarithmic space, the assumption of logarithmic relations inherent in the L1, L5, and AIC (in this instance) methods produced positively biased concentration and load estimates during high-flow conditions (fig. 7*A*) that resulted in positively biased load estimates over the entirety of the water year (fig. 7*B*). The inclusion of a streamflow-squared term via the L7 method reduced the rate at which total nitrogen concentrations increased across streamflows (fig. 7*A*), which better corresponded to sampled and observed values, and thus the L7 estimate was less biased than linear methods (fig. 7*B*). The general pattern of WRTDS estimates more closely matched sampled and observed values than INTERP and LOADEST-based methods (fig. 7); the resulting annual load estimate was only 11 percent greater than the observed annual load. Reflecting results presented previously, the adjustment of daily estimates based on sampled values by the AIC_COMP and WRTDS_K methods produced estimates that most closely approximated sampled concentrations and loads, particularly at high-flow conditions (figs. 7*A*, *B*). Resulting AIC_COMP and WRTDS_K annual total nitrogen load estimates nearly matched observed annual loads in this example (fig. 7*B*). The example in figure 7 highlights that (1) 18 samples per year (even with 6 samples specifically targeting higher streamflows) may not adequately represent actual loading conditions at sites with small drainages (like ROCK), (2) actual relations between total nitrogen concentrations and streamflow may not be adequately represented using prescribed relations with concentration and streamflow, and (3) adjusting these daily estimates based on samples tends to result in more accurate annual total nitrogen load estimates.

The relation among observed, sampled, and estimated nitrate plus nitrite concentrations and loads relative to streamflow conditions at ROCK in 2004 is illustrated in figure 8. Similar to total nitrogen loads (fig. 7), 80 percent of the annual load was transported during 45 days in which streamflows were greater than 55 ft^3/s (out of 335 days with observations). This example illustrates a common issue with the use of periodic sampling to estimate annual water-quality loads. Although the mean of sampled nitrate plus nitrite concentrations in figure 8*A* closely approximates observed concentrations throughout 2004, relatively few samples were collected during streamflows in which most of the annual load is transported. The only two samples recorded in 2004 that were greater than 55 ft^3/s were smaller than most observed values during the same streamflow conditions (fig. 8*A*), and thus estimation methods that rely exclusively on data from 2004 have the potential to underrepresent nitrate plus nitrite transport during the most influential conditions. However, methods that consider samples before 2004 are afforded a more accurate representation of observed values during high-flow conditions (figs. 8*A*, *B*).

As with figure 7, differences in the assumptions of estimation methods substantially affected the accuracy of daily and annual load estimates. Because nitrate plus nitrite concentrations were generally uncorrelated with streamflow greater than 65 ft^3/s (fig. 8*A*), INTERP produced estimates that were similar to sampled concentrations across streamflows; the resulting annual load estimate was biased only slightly low (−6.9 percent; fig. 8*B*). Although most LOADEST methods use data beyond the target year (with the exception of L1), assumptions about the relation among concentration and streamflow resulted in positively biased load estimates in figure 8. The assumption of linear (in logarithmic space) relations among concentration and streamflow within the L1 and L5 methods reasonably approximated nitrate plus nitrite concentrations during low streamflows but substantially overestimated concentrations at high streamflows (fig. 8*A*), resulting in annual load estimates that were more than double the observed load (fig. 8*B*). The assumption of quadratic relations among concentration and streamflow by L7 and AIC (in this example) more closely approximated observed concentrations during high-flow conditions than the L1 or L5 methods but also tended to overestimate observed concentrations, particularly for streamflows between 10 and 55 ft^3/s (fig. 8*A*). AIC_COMP, which adjusts AIC estimates based on sampled concentrations, better approximated sampled and observed loads; the resulting annual load estimate (+15 percent) was substantially more accurate than the AIC method alone (+63 percent; fig. 8*B*). Daily WRTDS and WRTDS_K estimates better characterized nitrate plus nitrite concentrations across low- and high-streamflow conditions (fig. 8*A*) because these methods (1) use more historical sample data and (2) use a weighted-regression approach that more heavily weights samples collected during similar streamflows, seasons, and times. WRTDS_K further improved upon WRTDS estimates by adjusting daily estimates to better match sampled values (figs. 8*A*, *B*).

Observed, sampled, and estimated total phosphorus loads relative to streamflow conditions at ROCK in 2004 are illustrated in figure 9. Total phosphorus concentrations are typically more positively correlated to streamflow conditions than total nitrogen or nitrate plus nitrite; in this example, 80 percent of the annual load was transported during only 28 days (out of 335 days with observations) with the highest streamflows

(greater than 100 ft^3/s). The transport of 80 percent of the annual load during less than 10 percent of possible days limits the ability of sampling to characterize loading conditions. HIFLOW sampling produced only one sample during high streamflows in 2004 and one additional sample from 2000 to 2003 (fig. 9).

The relative lack of high-flow samples in this example caused many methods to produce inaccurate annual load estimates. INTERP underrepresented concentrations during high-flow conditions, resulting in an annual load estimate that was less than half of the observed load (fig. 9*B*), which was similar to patterns observed for total phosphorus generally (table 6). The combination of relatively few observations during high streamflows and the assumption of linearity among the logarithm of concentration and streamflow conditions caused the L1 method to underestimate daily total phosphorus loads during high streamflows (−29 percent; fig. 9*B*), and thus for 2004 generally (−27 percent; fig. 9*B*). The consideration of samples from the previous 4 years and representation of the effects of season and time allowed the L5 method to better represent total phosphorus concentrations across streamflows; the annual L5 estimate was among the most accurate (−8.9 percent; fig. 9*B*) of the methods illustrated. The lack of high-flow samples and assumption of quadratic relations among the logarithm of concentration/streamflow conditions caused the L7 method to substantially overestimate total phosphorus concentrations/loads during high-flow conditions (59-percent error; fig. 9*B*), and thus for all of 2004 (42-percent error; fig. 9*B*). The AIC method used a cubic representation of the logarithm of concentration/streamflow relations in this example (along with seasonal and streamflow anomaly variables) and produced a relatively accurate representation of observed total phosphorus values across streamflow conditions (−6.8-percent error; fig. 9*B*). Consideration of additional historical data (fig. 2) along with a weighted-regression approach helped WRTDS produce among the most accurate representation of total phosphorus concentrations across streamflows (+8.4-percent error). The lack of sample data during high-flow conditions precluded the AIC_COMP and WRTDS_K (−17 percent and −13 percent of observed loads, respectively) methods from improving upon comparable methods (AIC and WRTDS) that did not adjust daily estimates based on sampled values. The frequent inability of AIC_COMP and WRTDS_K to improve upon the accuracy of annual AIC and WRTDS total phosphorus estimates is consistent with results observed generally (table 6).

The performance of selected methods relative to streamflow conditions for suspended sediment computed under the HIFLOW sampling strategy at POTO in 1978 is illustrated in figure 10. As with total phosphorus at ROCK, methods produced inaccurate load estimates, in part, because relatively few samples were available during days (35 of 365) in which streamflows transported 80 percent of the annual load. HIFLOW sampling in this example produced one sample during high-flow conditions during 1978 and four samples during these conditions from 1974 to 1977. As illustrated with total phosphorus in figure 9 (and broadly in table 7 and fig. 4), the interpolation of suspended sediment substantially underestimated concentrations and loads (−81 percent) during high streamflows and thus generally (−71 percent; fig. 10). Linear (L1 and L5) and quadratic (L7) representations (in logarithmic space) mischaracterized nonmonotonic observed relations between suspended sediment and streamflow conditions. Logarithmic relations among suspended sediment and streamflow used by the L1 and L5 methods underestimated suspended sediment during streamflows less than 2,000 ft^3/s, generally overestimated suspended sediment during streamflows between 2,000 and 20,000 ft^3/s, and underestimated suspended sediment during streamflows greater than 41,000 ft^3/s (fig. 10*A*). Because most of the annual load in this example was transported at streamflows greater than 41,000 ft^3/s, the L1 and L5 methods underestimated suspended-sediment loads in this example by −49 percent and −38 percent, respectively (fig. 10*B*). Quadratic relations between suspended sediment and streamflow (in logarithmic space) used by the L7 method overestimated observed concentrations and loads at streamflows beyond sampled values; the resulting L7 estimate in this example was more than double the observed load (182 percent; fig. 10*B*). As with total phosphorus, the AIC method used a cubic representation of suspended sediment and streamflow relations in logarithmic space (along with seasonal and streamflow-anomaly variables) in this example, and resulting estimates better mimicked observed concentrations as compared to the L1, L5, or L7 methods. However, the prescribed cubic relation between streamflow and streamflow still produced biased load estimates when forced to extrapolate beyond sampled streamflows and thus produced a positively biased annual load estimate (+38 percent) for all of 1978 (fig. 10*B*). As with the total phosphorus example (fig. 9), the consideration of data beyond the 5-year sampling window (fig. 2) along with a weighted-regression approach allowed WRTDS to better represent observed relations between suspended sediment and streamflow throughout streamflow conditions in the POTO example; the resulting annual load estimate was the most accurate among all methods (2.3-percent error). Also similar to the total phosphorus example (fig. 9), the lack of sample collection during high streamflows precluded methods that adjust daily estimates based on sample data (AIC_COMP, 45-percent error; WRTDS_K, −13-percent error) from improving upon AIC and WRTDS estimates in this example. Although AIC_COMP and WRTDS_K did not improve upon AIC and WRTDS_K results in this example, it is important to note that they did produce more accurate results across suspended-sediment sites generally (table 7).

Examples in figures 7–10 are included to provide context to aggregated results presented in previous sections. Constituents that increase in concentration during high-flow conditions, such as total phosphorus and suspended sediment, are transported primarily during high-flow events that encompass a relatively small proportion of the year. For these constituents, even purposeful high-flow sampling strategies can mischaracterize true relations among concentration and

streamflow; this is especially true for methods that use data for the target year only. The assumption of defined, linear, or quadratic relations among streamflow and concentration inherent in the L1, L5, and L7 methods can misrepresent sites/constituents with more complex water-quality transport processes. The assumption of static relations among concentration and streamflow is especially likely to result in biased load computations when methods are forced to extrapolate beyond sampled values. The AIC method also is subject to this limitation; however, cubic relations used in most of the examples provided an improved representation of observed relations among concentration and streamflow. WRTDS produced among the most accurate loads in these examples because it considered samples beyond the previous 5 years and because relations among concentration and streamflow at the highest streamflow conditions were modeled using samples collected during similar conditions. Adjustments to daily estimates based on sampled values were particularly effective for total nitrogen and nitrate plus nitrite examples in which target-year samples were similar to observed values for a given streamflow condition. Adjustments used as part of the WRTDS_K and AIC_COMP methods were less effective for total phosphorus and suspended-sediment examples because few samples were available during high-flow conditions that produced 80 percent of the annual load. However, it is important to note that WRTDS_K and AIC_COMP produced more estimates within thresholds than the WRTDS and AIC methods for total phosphorus and suspended sediment generally (tables 6–7; fig. 4).

Causes of Error among Estimation Methods

Previously presented results indicate that the WRTDS_K and, to a lesser degree, AIC_COMP methods are most likely to produce accurate annual load estimates among multiple water-quality constituents. However, these methods still have the potential to produce biased estimates, and thus it is desirable to better understand factors that affect the computation of accurate (or inaccurate) annual water-quality loads. A regression-tree approach (appendix 7) was used to characterize if metrics computed from continuous discharge and periodic water-quality sampling records could predict the accuracy of WRTDS_K, AIC_COMP, and AIC-computed annual water-quality load estimates. The AIC method was evaluated in addition to the WRTDS_K and AIC_COMP methods to evaluate the ability of regression trees to predict the accuracy of methods that do not adjust daily estimates based on sampled values. Streamflow and sampling record metrics evaluated include measures of the amount of base streamflow in the daily streamflow record, the variability of discrete concentration and load observations, the mean of the daily streamflow record, the number of discrete water-quality observations, measures of how well samples represent peak-flow conditions, the correlation among discrete water-quality concentration and streamflow, and measures of the slope of discrete water-quality concentration and streamflow.

Regression trees had somewhat limited success in predicting if load estimates fell within predefined accuracy thresholds (appendix 7); however, the manner in which sampling record characteristics predicted load-estimate accuracy was similar among methods and water-quality constituents. More variable sampled concentrations and loads, more runoff, higher slopes among concentration and streamflow values, and less representation of peak-flow conditions were generally predictive of less accurate load estimates. The consideration of alternate sampling record characteristics and (or) the use of different techniques may offer an improved ability to identify biased estimates. A more complete description of the methods and results of this analysis is provided in appendix 7.

Discussion

The impetus of this study was to expand upon an evaluation of methods for computing decadal results presented in Lee and others (2016) to consider annual loads. Patterns observed for decadal loads among water-quality constituents, sampling strategies, sampling sites, and estimation methods were similar to results presented for an annual time step. As with decadal loads, estimation method accuracy generally decreased for annual total nitrogen, nitrate plus nitrite, total phosphorus, and suspended-sediment loads (chloride was not assessed for decadal loads). Among sampling strategies with the same sampling frequency, purposeful collection of samples during high-flow conditions generally resulted in the most accurate annual and decadal-load estimates. Annual and decadal loads also were more difficult to estimate at sites with smaller drainages and more streamflow conditions. Methods that assume linear or quadratic relations among the logarithm of concentration/streamflow conditions, such as L1, L5, L7, and LAICO, frequently produced less accurate annual and decadal loads compared to methods that included cubic transformations of streamflow, used more flexible relations among concentration and discharge (WRTDS), or adjusted daily load estimates based on departures from observed values (WRTDS_K and AIC_COMP). For total nitrogen and nitrate plus nitrite, interpolation and ratio estimation produced among the most accurate estimates for annual and decadal loads; however, these methods were among the least accurate for computing annual chloride, total phosphorus, or suspended-sediment loads. Interpolation and ratio estimation were likely more accurate when computing total nitrogen and nitrate plus nitrite loads because concentrations of these constituents are typically less correlated with streamflow conditions, and thus methods that do not specify a specific relation among concentration and streamflow, such as interpolation or ratio estimation, are more likely to produce accurate load estimates.

All methods computed annual loads within predefined accuracy thresholds much less frequently than for decadal loads. For the same method and constituent, and among similar sampling strategies, annual estimates were within

±10 percent or ±20 percent of observed loads among water-quality constituents 21 to 64 percent as often as decadal loads (percentages of accuracy vary because different thresholds were compared among water-quality constituents in Lee and others [2016]). For example, decadal suspended-sediment loads computed using WRTDS were within ±20 percent of observed loads for 81 percent of cases (see table 2 in Lee and others [2016] for more details), whereas an average of 26 percent of WRTDS loads computed under the NWQN, HIFLOW, HIFLOWE, BIWEEK, and MONTH sampling strategies were within ±20 percent of observed annual suspended-sediment loads (table 7). The substantial reduction in the accuracy of estimation methods for computing annual loads should be noted by researchers assessing water-quality effects on receiving waters, quantifying surface water-quality trends, and modeling the effects of landscape practices on water-quality conditions.

Although there were similarities in the performance of estimation methods at annual and decadal time steps, this study expanded the number of estimation methods considered and offered additional analysis of factors affecting annual load-estimate accuracy. In contrast to decadal-load findings, in which no one method was identified as the most accurate across water-quality constituents, WRTDS_K (which was not considered in the decadal study) was determined to generally produce the most accurate annual loads among multiple water-quality constituents. WRTDS_K was often more accurate than other methods because (1) it incorporates more historical water-quality data, thus reducing the potential that limited sampling will inaccurately characterize observed daily concentrations and loads; (2) weighted regressions allow WRTDS_K to account for nonstationarity in relations among concentration, streamflow, season, and time; and (3) WRTDS_K adjusts daily load estimates based on departures from measured values, which often substantially improved the accuracy of annual load estimates. The WRTDS_K method was not available for evaluation by Lee and others (2016); however, based on improvements in accuracy observed when using Kalman filters for computing for annual loads via the FLUXMASTER program (Lee and others, 2016), WRTDS_K also would likely improve the accuracy of decadal-load estimates. It is important to note that difficulties associated with adequately sampling high-flow conditions, especially when computing total phosphorus or suspended-sediment loads, will limit the ability of any methods to improve load-estimate accuracy.

Another contrast to the results presented by Lee and others (2016) is that ratio estimation, which was among the best performing methods for decadal loads, was often among the least accurate methods for computing annual loads. Ratio estimators that used data from the target year only (RATIO_T and RATIO_F1) were among the six most accurate methods for computing total nitrogen and nitrate plus nitrite annual loads (but less accurate than the WRTDS_K or AIC_COMP methods; tables 4–5) but were among the worst performing methods for computing total phosphorus and suspended-sediment annual loads (tables 6–7). This result contrasts with those presented in Lee and others (2016), in which ratio estimation produced the second most estimates within ±10 percent of observed nitrate plus nitrite loads and produced the most estimates within ±20 percent of observed total phosphorus and suspended-sediment loads (table 2 in Lee and others, 2016). Differences in the performance of ratio estimation between decadal and annual loads are likely related to the number of samples considered. Under monthly sampling, ratio estimators use 120 samples over a decade, which allows for a better characterization of actual concentration/flow ratios than the 12 samples considered at an annual time step. The poor performance of the ratio estimators in this study is primarily attributed to the relatively small sample size used for each stratum, especially for methods that are restricted to using data from only a single year.

The extensive dataset compiled for this study presented an opportunity to test if sampling record characteristics could identify if annual load estimates are likely to be accurate. Although the regression-tree analysis (detailed in appendix 7) failed to characterize the cause of estimation method accuracy in most cases, larger slopes among the logarithm of concentration and streamflow, more variability in sampled concentrations and (or) loads, more runoff, and less representation of peak observed streamflow conditions generally led to reduced load-estimate accuracy. These findings, along with examples that illustrated method performance relative to daily streamflow conditions, indicated that (1) even purposeful high-flow sampling may not adequately characterize actual water-quality transport patterns when most annual loads are transported during a few days; (2) relations among concentration and streamflow are often complex and not adequately specified via linear or quadratic relations; and (3) although localized adjustments of daily load estimates based on sampled results improve annual estimates generally, adjustments do not necessarily improve estimates when relatively few samples are collected during periods in which most loads are transported.

The findings in this study have several implications for practitioners computing water-quality loads. First, the collection of 26 samples per year generally improved the accuracy of annual load estimates as compared to the collection of 18, 12, or 6 samples per year, regardless of sampling strategy. Among sampling strategies, the purposeful collection of samples at high-flow conditions generally improved load-estimate accuracy relative to seasonally weighted sampling, regardless of the time of year in which high-flow samples were collected. Second, for chloride or total nitrogen, one can expect to compute relatively accurate (±20 percent of observed) annual loads with many estimation methods and sampling strategies. However, the selection of sampling strategy and estimation method becomes more important when computing nitrate plus nitrite, total phosphorus, or suspended-sediment loads, especially at sampling sites with small drainages and (or) variable streamflow/loading conditions. For suspended sediment in particular, most estimation methods produced estimates outside of ±20 percent of observed loads at sites with small drainages and (or) variable streamflow conditions. When estimating

loads in these cases, it may be more appropriate to investigate alternative approaches, such as using continuous water-quality sensors to serve as surrogates for water-quality concentrations (Robertson and others, 2018). Finally, the results presented herein indicate that WRTDS_K is likely the best method for practitioners who desire a single method to estimate loads at multiple sites or for multiple water-quality constituents. In addition, the underlying WRTDS method (on which the WRTDS_K estimates are based) includes a robust "flow-normalization" approach that allows researchers to assess how water-quality loads change independent of variation in streamflow conditions. This ability, in combination with error estimation, gives practitioners the ability to quantify the magnitude and certainty of water-quality trends over various time frames such as a decade or multiple decades. Relatively accurate quantification of annual loads and the capacity of trend analysis make WRTDS_K a valuable tool for researchers who want to characterize if changes in upstream basins are affecting downstream water-quality concentrations or loads.

The NWQN computed loads using the LAICO method from 1963 to 2012 and a modified version of the AIC/PVAL methods from 2013 to present (2019; Lee and others, 2017a). Although Lee and others (2017a) detailed plans to use WRTDS to compute loads at NWQN sites beginning in 2017, WRTDS was used at only two sites to characterize changes in water-quality loading to the Gulf of Mexico (Lee and others, 2017b). WRTDS was not used at all NWQN sites because of pending updates to the method, including WRTDS_K and improvements to streamflow-normalization processes (Choquette and others, 2019). Based on conclusions from this study, WRTDS_K is planned to be used to compute NWQN loads along with the adapted-LOADEST method described in Lee and others (2017a) and evaluated in appendix 6. The adapted-LOADEST method will continue to be used to compute loads at NWQN sites to maintain consistency with historical estimation procedures and because results presented in this study indicate that this method is likely to produce accurate water-quality loads at large river sites that have less variable loading conditions.

Although this study contributed new information regarding load estimation at an annual time step, multiple questions remain unresolved. Reliable methods for computing water-quality loads at headwater stream sites, especially for computing total phosphorus and suspended-sediment loads, still need to be identified. Further investigation is needed to identify values used for default WRTDS_K settings, such as the lag-1 autocorrelation coefficient, for improving the accuracy of load estimates for different water-quality constituents and sampling sites. Also, relatively few studies have evaluated the degree to which sensor technologies such as nitrate, dissolved phosphorus, or turbidity can improve load-estimate accuracy, especially when considering the likelihood of sensor fouling and maintenance issues (Robertson and others, 2018). Although most methods described herein offer methods to compute the uncertainty of load estimates, more study is needed to characterize the accuracy of these estimates. Finally, although this report focused on annual loads, many water-quality effects, such as algal blooms or hypoxia, may be better explained by seasonal or daily constituent loads; future work is needed to assess the accuracy of methods for computing these estimates.

Summary and Conclusions

This study evaluates methods for computing annual water-quality loads, specifically with respect to methods currently (2019) used for sites in the U.S. Geological Survey National Water Quality Network. Near-daily datasets of chloride, total nitrogen, nitrate plus nitrite, total phosphorus, and suspended sediment were subset to determine the accuracy of various load-estimation methods, including linear interpolation, ratio estimators, LOADEST-based regression methods, and weighted regression. Methods were evaluated for different sampling strategies, among different water-quality constituents, and at different sampling sites.

Estimation methods were less accurate when computing loads at annual rather than decadal time steps. Depending on the water-quality constituent, annual loads were within comparable accuracy thresholds 21 to 64 percent of the time relative to decadal loads. The frequency and methods by which water-quality samples were collected and the water-quality constituents that were estimated had important implications for the accuracy of annual load estimates. The collection of 26 samples per year improved the accuracy of annual load estimates as compared to the collection of 18, 12, or 6 samples per year, regardless of sampling strategy. Among sampling strategies, the purposeful collection of samples at high-flow conditions generally improved load-estimate accuracy relative to seasonally weighted sampling. Among water-quality constituents, relatively accurate (±20 percent of observed loads) chloride and total nitrogen loads were computed by many estimation methods and sampling strategies. However, the choice of sampling strategy and estimation method was more important for computing nitrate plus nitrite, total phosphorus, and suspended-sediment loads, especially at sampling sites with small drainages and (or) variable streamflow/loading conditions.

In terms of specific estimation methods, the Weighted Regressions on Time, Discharge, and Season method with Kalman filtering generally produced the most accurate annual load estimates among sampling sites and water-quality constituents. Linear interpolation and ratio estimators that only used samples from the year being estimated were among the most likely to produce accurate total nitrogen and nitrate plus nitrite loads but were among the least likely to produce accurate total phosphorus and suspended-sediment loads. LOADEST-based methods that specified linear or quadratic relations among concentration and streamflow (in logarithmic space) were generally among the least accurate methods, although the LOADEST-based methods that considered cubic relations among the logarithm of concentration and streamflow were

more likely to produce accurate loads. Methods that adjusted daily estimates computed from regression (or weighted-regression) methods based on departures from sampled values, such as the Weighted Regressions on Time, Discharge, and Season method with Kalman filtering and the composite method, were more likely to produce accurate estimates generally, but especially when computing total nitrogen, nitrate plus nitrite, and suspended-sediment loads.

Based on the findings from this report, the U.S. Geological Survey plans to continue to publish water-quality loads using LOADEST-based methods that consider multiple transformations of National Water Quality Network streamflow, as well as season, time, and variables indicative of historical streamflow conditions, to preserve historical records used by stakeholders. However, the U.S. Geological Survey also plans to publish annual load estimates using the Weighted Regressions on Time, Discharge, and Season method with Kalman filtering because these estimates have been determined to be the most likely to be accurate for a given site, constituent, and water year.

References Cited

Akaike, H., 1974, A new look at the statistical model identification: IEEE Transactions on Automatic Control, v. 19, no. 6, p. 716–723. [Also available at https://doi.org/10.1109/TAC.1974.1100705.]

Appling, A.P., Leon, M.C., and McDowell, W.H., 2015, Reducing bias and quantifying uncertainty in watershed flux estimates—The R package loadflex: Ecosphere, v. 6, no. 12, p. 1–25. [Also available at https://doi.org/10.1890/ES14-00517.1.]

Aulenbach, B.T., and Hooper, R.P., 2006, The composite method—An improved method for stream-water solute load estimation: Hydrological Processes, v. 20, no. 14, p. 3029–3047. [Also available at https://doi.org/10.1002/hyp.6147.]

Baun, K., 1982, Alternative methods of estimating pollutant loads in flowing water: Wisconsin Department of Natural Resources Technical Bulletin 133, p. 1–16.

Beale, E.M.L., 1962, Some uses of computers in operational research: Industrielle Organisaton, v. 31, p. 51–52.

Choquette, A.F., Hirsch, R.M., Murphy, J.C., Johnson, L.T., and Confesor, R.B., Jr., 2019, Tracking changes in nutrient delivery to western Lake Erie—Approaches to compensate for variability and trends in streamflow: Journal of Great Lakes Research, v. 45, no. 1, p. 21–39. [Also available at https://doi.org/10.1016/j.jglr.2018.11.012.]

Cochran, W.G., 1977, Sampling techniques, 3d ed.: New York, Wiley, 428 p.

Cohn, T.A., Caulder, D.L., Gilroy, E.J., Zynjuk, L.D., and 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 Resources Research, v. 28, no. 9, p. 2353–2363. [Also available at https://doi.org/10.1029/92WR01008.]

Cohn, T.A., DeLong, L.L., Gilroy, E.J., Hirsch, R.M., and Wells, D.K., 1989, Estimating constituent loads: Water Resources Research, v. 25, no. 5, p. 937–942. [Also available at https://doi.org/10.1029/WR025i005p00937.]

Deacon, J.R., Lee, C.J., Toccalino, P.L., Warren, M.P., Baker, N.T., Crawford, C.G., Gilliom, R.G., and Woodside, M.D., 2015, Tracking water-quality of the Nation's rivers and streams: U.S. Geological Survey web page, accessed July 2017 at https://doi.org/10.5066/F70G3H51.

Dolan, D.M., Yui, A.K., and Geist, R.D., 1981, Evaluation of river load estimation for total phosphorus: Journal of Great Lakes Research, v. 7, no. 3, p. 207–214. [Also available at https://doi.org/10.1016/S0380-1330(81)72047-1.]

Ferguson, R.I., 1986, River loads underestimated by rating curves: Water Resources Research, v. 22, no. 1, p. 74–76. [Also available at https://doi.org/10.1029/WR022i001p00074.]

Heidelberg University, 2005, Tributary data download: Heidelberg University web page, accessed March 2016 at https://www.heidelberg.edu/tributary-data-download.

Hirsch, R.M., 2014, Large biases in regression-based constituent load estimates—Causes and diagnostic tools: Journal of the American Water Resources Association, v. 50, no. 6, p. 1401–1424. [Also available at https://doi.org/10.1111/jawr.12195.]

Hirsch, R.M., Archfield, S.A., and De Cicco, L.A., 2015, A bootstrap method for estimating uncertainty of water quality trends: Environmental Modelling & Software, v. 73, p. 148–166. [Also available at https://doi.org/10.1016/j.envsoft.2015.07.017.]

Hirsch, R.M., Moyer, D.L., and Archfield, S.A., 2010, Weighted Regressions on Time, Discharge, and Season (WRTDS), with an application to Chesapeake Bay River inputs: Journal of the American Water Resources Association, v. 46, no. 5, p. 857–880. [Also available at https://doi.org/10.1111/j.1752-1688.2010.00482.x.]

Kalman, R.E., 1960, A new approach to linear filtering and prediction problems: Journal of Basic Engineering, v. 82, no. 1, p. 35–45. [Also available at https://doi.org/10.1115/1.3662552.]

Lee, C.J., 2019, Supplementary data used to evaluate methods for computing annual water-quality loads, 1948–2016: U.S. Geological Survey data release, https://doi.org/10.5066/P9BK91LN.

Lee, C.J., Henderson, R.J., and Deacon, J.D., 2017b, Nutrient loading for the Mississippi River Basin and subbasins: U.S. Geological Survey web page, accessed September 4, 2017, at https://nrtwq.usgs.gov/mississippi_loads/#/.

Lee, C.J., Hirsch, R.M., Schwarz, G.E., Holtschlag, D.J., Preston, S.D., Crawford, C.G., and Vecchia, A.V., 2016, An evaluation of methods for estimating decadal stream loads: Journal of Hydrology (Amsterdam), v. 542, p. 185–203. [Also available at https://doi.org/10.1016/j.jhydrol.2016.08.059.]

Lee, C.J., Murphy, J.C., Crawford, C.G., and Deacon, J.R., 2017a, Methods for computing water-quality loads at sites in the U.S. Geological Survey National Water Quality Network: U.S. Geological Survey Open-File Report 2017–1120, 20 p., accessed July 2018 at https://doi.org/10.3133/ofr20171120.

Maccoux, M.J., Dove, A., Backus, S.M., and Dolan, D.M., 2016, Total and soluble reactive phosphorus loadings to Lake Erie—A detailed accounting by year, basin, country, and tributary: Journal of Great Lakes Research, v. 42, no. 6, p. 1151–1165. [Also available at https://doi.org/10.1016/j.jglr.2016.08.005.]

Moyer, D.L., Hirsch, R.M., and 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, 118 p., accessed February ,2016 at https://doi.org/10.3133/sir20125244.

R Core Team, 2017, R—A language and environment for statistical computing: Vienna, Austria, R Foundation for Statistical Computing, 3693 p. [Also available at https://cran.r-project.org/.]

Richards, R.P., 1998, Estimation of pollutant loads in rivers and streams—A guidance document for NPS programs: Tiffin, Ohio, U.S. Environmental Protection Agency, 134 p., accessed June 2016 at http://abca.iwebsmart.net/downloads/Richards-1998.pdf.

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., and Scavia, D., 2012, Discussion—Nutrient inputs to the Laurentian Great Lakes by source and watershed estimated using SPARROW watershed models: Journal of the American Water Resources Association, v. 49, no. 3, p. 715–724. [Also available at https://doi.org/10.1111/jawr.12006.]

Robertson, D.M., Hubbard, L.E., Lorenz, D.L., and Sullivan, D.J., 2018, A surrogate regression approach for computing continuous loads for the tributary nutrient and sediment monitoring program on the Great Lakes: Journal of Great Lakes Research, v. 44, no. 1, p. 26–42. [Also available at https://doi.org/10.1016/j.jglr.2017.10.003.]

Runkel, R.L., Crawford, C.G., and 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, chap. A5, 69 p., accessed June 2016 at https://doi.org/10.3133/tm4A5.

Ryberg, K.R., and 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, 8 p., accessed June 2016 at https://doi.org/10.3133/ofr20121168.

Stenback, G.A., Crumpton, W.A., Schilling, K.E., and Helmers, M.J., 2011, Rating curve estimation of nutrient loads in Iowa rivers: Journal of Hydrology (Amsterdam), v. 396, no. 1–2, p. 158–169. [Also available at https://doi.org/10.1016/j.jhydrol.2010.11.006.]

Tin, M., 1965, Comparison of some ratio estimators: Journal of the American Statistical Association, v. 60, no. 309, p. 294–307. [Also available at https://doi.org/10.1080/01621459.1965.10480792.]

U.S. Geological Survey, 2017, USGS water data for the Nation: U.S. Geological Survey National Water Information System database, accessed March 16, 2017, at https://doi.org/10.5066/F7P55KJN.

Wood, S.N., 2006, Generalized additive models—An introduction with R: Boca Raton, Fla., Chapman and Hall/CRC, 410 p.

Appendixes 1–7

Appendix 1. Description of Weighted Regressions on Time, Discharge, and Season Method with Kalman Filtering

The Weighted Regressions on Time, Discharge, and Season method with Kalman filtering (WRTDS_K) is a variation on the Weighted Regressions on Time, Discharge, and Season (WRTDS) method for estimating concentration as a function of time, discharge, and season. When placed in a time-series context, with the time step set to one observation per day, the WRTDS model can be expressed as follows:

$$\ln(c_i) = \beta_{i0} + \beta_{i1}\ln(Q_i) + \beta_{i2}T_i + \beta_{i3}\sin(2\pi T_i) + \beta_{i4}\cos(2\pi T_i) + \sigma_i z_i \quad (1.1)$$

where

c_i is concentration on day i, in milligrams per liter;

$\beta_{i0}, \beta_{i1}, \beta_{i2}, \beta_{i3}, \beta_{i4}$ are all fitted coefficients of the model that vary smoothly over the model domain (that is, over time and discharge). Each of them has a specific value for each day in the record, based on the T_i and Q_i values for that day;

Q_i is the mean daily discharge on day i, in cubic meters per second;

T_i is time expressed as decimal year;

σ_i is the fitted value of the conditional standard deviation of the error component of the model. It also varies smoothly over the model domain as a function of T_i and Q_i; and

z_i is the standardized model residual on day i (standard deviation = 1).

The model is fit in the standard manner of WRTDS models, using the EGRET software package. Samples used by the model are assumed to be available every few weeks to every few months and may be collected at irregular time intervals. Once the model has been fit to sampled data, we can calculate residuals (r_i) for days in which we have concentration measurements. We compute the residuals as follows:

$$r_i = \ln(c_i) - (\beta_{i0} + \beta_{i1}\ln(Q_i) + \beta_{i2}T_i + \beta_{i3}\sin(2\pi T_i) + \beta_{i4}\cos(2\pi T_i)) \quad (1.2)$$

Residuals are the error in the model predictions, expressed in logarithmic space (that is, they are the observed logarithmic concentration minus the predicted logarithmic concentration). We can take one additional step and standardize these residuals by dividing by the standard deviation appropriate to that day. These standardized residuals we will call z_i.

They are computed as

$$z_i = \frac{r_i}{\sigma_i} \quad (1.3)$$

The first point where WRTDS_K departs from the standard WRTDS method is what it uses to estimate the expected value of concentration on a day for which there is a sample value ($E[c_i]$). In the standard WRTDS method, the estimate of concentration on sampled days is not the observed value but is actually the unbiased estimate from the WRTDS model. This unbiased estimate is

$$E[c_i] = \exp\left\{\beta_{i0} + \beta_{i1}\ln(Q_i) + \beta_{i2}T_i + \beta_{i3}\sin(2\pi T_i) + \beta_{i4}\cos(2\pi T_i) + \frac{\sigma_i^2}{2}\right\} \quad (1.4)$$

The final term in this equation is the bias correction factor that is required to convert the modeled natural logarithm of concentration to arithmetic space. This bias correction factor is approximately correct when the errors in logarithmic space are normal, σ_i is relatively small, and the sample size is large (say greater than 50). In WRTDS_K, the estimate for days with samples is the sample value rather than the expected value using the model (eq. 4). Clearly, we will improve our overall accuracy if we use data rather than estimates on those days when we have data.

The other way WRTDS_K departs from WRTDS is how the method estimates concentrations on days where there is no sample value. In WRTDS_K, the estimate for a given day makes use of the measured data from the most recent preceding measurement and the next succeeding measurement. Based on experience, we know that the standardized residuals likely have a good deal of serial correlation. In WRTDS_K, as currently implemented, we assume that the serial correlation structure of the z_i values is autoregressive lag 1, and we further assume that the autoregressive lag-1 correlation coefficient, probability (p) equals 0.95. Further refinement of this method will probably lead to an approach to accurately estimate p from the irregularly spaced data, but for now, this is the approach being proposed. The results are not highly sensitive to the choice of p as long as it is in a range from about 0.8 to 0.95.

Estimates for unsampled days are computed by first dividing the record into sets of consecutive unsampled days. For each of these periods, the day of the last observation before the unsampled period is considered day 1, and the day of the first observation after the unsampled period is day n; thus, there are $n-2$ observations in the unsampled period that we would like to estimate. Because of the complexity of the process (for example, logarithmic transformations, time varying model coefficients, and time varying variances), we estimate the expected value for each of the $n-2$ missing values using a Monte Carlo simulation (50 replicates were used in this study). For each replicate, the method generates a time series of the $n-2$ values for the unsampled period. For any given replicate of data for the unsampled period, these

values are $c_2, c_3, c_4 \ldots, c_{n-3}, c_{n-2}, c_{n-1}$. Computing estimates for unsampled days depends on knowing c_1 and c_n and knowing all the parameters of equation 1.1, the estimate of σ for each day during the unsampled period, and the distributional form and correlation structure of the error term, ε_i. The generation of a single replicate (call it replicate m) of these $n-2$ values of concentration is completed as follows:

1. Generate $n-2$ values of e_k, which are independent standard normal random variables (mean 0, variance 1).

2. Conditioned on the estimated values of the standardized residuals z_1 and z_n (determined from the data, the fitted WRTDS model, and eqs. 1.2 and 1.3), the remaining $n-2$ z_k values are generated based on the recursive relation of an autoregressive lag-1 process:

$$z_{k+1} = \rho z_k + \sqrt{1-\rho^2}\, e_k \text{ for } (1 < k < n) \tag{1.5}$$

Note that unlike the usual way that an autoregressive lag-1 process is generated, the generating process used here is conditioned on two known (nonrandom) values, one (z_1) that represents the day before the unsampled period, and the other (z_n) that represents the day after the unsampled period.

3. This process is repeated for every unsampled period in the record. This generating process is designed so that the entire time series of values of z (including the values calculated from the data and all the generated values in between them) have an autoregressive lag-1 correlation coefficient with an expected value of p.

4. This time series of z values is then transformed to a set of concentration (c) values using the fitted WRTDS model using equations 1.6 and 1.7:

$$c_i = \exp\{\hat{y}_i + \sigma_i z_i\} \tag{1.6}$$

where

$$\hat{y} = \beta_{i0} + \beta_{i1}\ln(Q_i) + \beta_{i2}T_i + \beta i_3 \sin(2\pi T_i) + \beta_{i4}\cos(2\pi T_i) \tag{1.7}$$

Note the difference between equations 1.4 and 1.6. In equation 1.4, the quantity being estimated is the expected value of c_i, but in equation 1.6, the quantity being estimated is a single realization of c_i. The bias correction term in equation 1.4 is not used here because we are not estimating a mean value; rather, we are estimating a single value. This process (steps 1 through 4) is repeated 50 times, and the expected value of concentration for each day (i) is the mean of the 50 replicates of c_i. This expected value of concentration for day i can be called $\overline{c}_i$. Note that in the special case of a measured day, all 50 replicates of c_i are equal to the observed value for that day. In all other cases, the 50 replicates of c_i include some variability.

Thus, the estimated mean load (in kilograms per day) for any given year is

$$\sum_{j=1}^{365} \overline{c}_j Q_j 86.4 \tag{1.8}$$

where

j is the day index for the days of the given year (rather than a single index of days from the start of the record to the end). Note that Q_j is in cubic meters per second and 86.4 is a unit conversion factor.

In general, when the sampling is sparse, such as in records with bimonthly sampling, the WRTDS_K approach will produce estimates that are similar to those determined in the original WRTDS method. However, when sampling is relatively frequent, such as at weekly intervals, the WRTDS_K approach will produce estimates that can be quite different from those determined in the original method. That is because there are many measured data values that WRTDS_K will use in place of an estimated value and because the serial dependence of the data at short lags (say 1 to 7 days) can have a strong effect on the estimates.

Appendix 2. Tables Indicating the Percentage of Annual Load Estimates within 10 Percent of Observed Loads among Methods and Sampling Strategies

Table 2.1. Percentage of annual chloride load estimates within plus or minus 10 percent of observed loads.

[Rows are sorted by methods with the most to least estimates within criteria in aggregate. Data are shaded in ranges of 80–100 percent, 70–79 percent, 60–69 percent, 50–59 percent, 40–49 percent, 30–39 percent, and 20–29 percent. BIWEEK, biweekly sampling; HIFLOW, high-flow sampling; HIFLOWE, high-flow early sampling; NWQN, National Water Quality Network sampling; MONTH, monthly sampling; BIMONTH, bimonthly sampling]

Method	BIWEEK, in percent	HIFLOW, in percent	HIFLOWE, in percent	NWQN, in percent	MONTH, in percent	BIMONTH, in percent
AIC_COMP	89	81	85	80	74	57
WRTDS_K	85	78	80	78	73	64
AIC	82	78	79	75	72	57
PVAL	81	76	78	73	70	58
L7	73	70	72	69	67	55
LAICO	73	68	71	67	64	54
L5	63	63	62	60	60	55
L1	71	69	60	62	57	49
WRTDS	62	60	61	59	58	55
RATIO_F5	52	47	47	46	47	43
RATIO_T	56	47	47	46	41	30
RATIO_F1	56	45	45	44	38	30
INTERP	38	32	32	28	28	21

Table 2.2. Percentage of annual total nitrogen load estimates within plus or minus 10 percent of observed loads.

[Rows are sorted by methods with the most to least estimates within criteria in aggregate. Data are shaded in ranges of 70–79 percent, 60–69 percent, 50–59 percent, 40–49 percent, 30–39 percent, and 20–29 percent. BIWEEK, biweekly sampling; HIFLOW, high-flow sampling; HIFLOWE, high-flow early sampling; NWQN, National Water Quality Network sampling; MONTH, monthly sampling; BIMONTH, bimonthly sampling]

Method	BIWEEK, in percent	HIFLOW, in percent	HIFLOWE, in percent	NWQN, in percent	MONTH, in percent	BIMONTH, in percent
WRTDS_K	78	73	71	67	63	49
AIC_COMP	71	60	59	62	53	34
RATIO_T	66	53	51	54	43	34
RATIO_F1	64	55	52	47	45	34
INTERP	60	48	48	50	40	34
AIC	51	48	49	45	45	32
PVAL	49	46	47	43	41	32
L7	46	44	45	41	42	36
WRTDS	42	40	40	42	41	37
LAICO	43	42	43	39	39	32
RATIO_F5	33	34	33	38	32	34
L1	30	30	30	27	29	27
L5	30	30	30	27	29	27

Table 2.3. Percentage of annual nitrate plus nitrite load estimates within plus or minus 10 percent of observed loads.

[Rows are sorted by methods with the most to least estimates within criteria in aggregate. Data are shaded in ranges of 70–79 percent, 60–69 percent, 50–59 percent, 40–49 percent, 30–39 percent, 20–29 percent, and 0–19 percent. BIWEEK, biweekly sampling; HIFLOW, high-flow sampling; HIFLOWE, high-flow early sampling; NWQN, National Water Quality Network sampling; MONTH, monthly sampling; BIMONTH, bimonthly sampling]

Method	BIWEEK, in percent	HIFLOW, in percent	HIFLOWE, in percent	NWQN, in percent	MONTH, in percent	BIMONTH, in percent
WRTDS_K	72	64	65	60	56	40
INTERP	71	55	57	56	46	40
RATIO_T	69	54	55	54	45	34
AIC_COMP	63	50	51	51	44	27
RATIO_F1	59	49	49	45	41	34
WRTDS	37	36	37	40	35	33
RATIO_F5	29	28	27	31	27	29
AIC	30	28	35	24	29	20
PVAL	30	28	35	24	29	20
L7	27	29	33	22	27	20
LAICO	27	27	33	24	27	20
L1	15	17	16	15	18	18
L5	15	16	18	11	15	16

Table 2.4. Percentage of annual total phosphorus estimates within plus or minus 10 percent of observed loads.

[Rows are sorted by methods with the most to least estimates within criteria in aggregate. Data are shaded in ranges of 50–59 percent, 40–49 percent, 30–39 percent, 20–29 percent, and 0–19 percent. BIWEEK, biweekly sampling; HIFLOW, high-flow sampling; HIFLOWE, high-flow early sampling; NWQN, National Water Quality Network sampling; MONTH, monthly sampling; BIMONTH, bimonthly sampling]

Method	BIWEEK, in percent	HIFLOW, in percent	HIFLOWE, in percent	NWQN, in percent	MONTH, in percent	BIMONTH, in percent
WRTDS_K	56	45	47	44	41	32
WRTDS	49	42	41	43	39	32
AIC	44	42	42	40	35	26
RATIO_F5	45	40	38	41	36	29
AIC_COMP	48	38	43	40	35	25
PVAL	43	41	40	40	35	26
L7	36	37	37	36	31	28
LAICO	36	37	34	36	31	28
L5	36	32	30	32	32	27
L1	30	26	27	29	24	19
RATIO_F1	31	25	25	28	21	14
RATIO_T	26	18	17	19	16	14
INTERP	17	14	13	11	11	9

Table 2.5. Percentage of annual suspended-sediment load estimates within plus or minus 10 percent of observed loads.

[Rows are sorted by methods with the most to least estimates within criteria in aggregate. Data are shaded in ranges of 40–49 percent, 30–39 percent, 20–29 percent, and 0–19 percent. BIWEEK, biweekly sampling; HIFLOW, high-flow sampling; HIFLOWE, high-flow early sampling; NWQN, National Water Quality Network sampling; MONTH, monthly sampling; BIMONTH, bimonthly sampling]

Method	BIWEEK, in percent	HIFLOW, in percent	HIFLOWE, in percent	NWQN, in percent	MONTH, in percent	BIMONTH, in percent
WRTDS_K	43	44	43	38	31	22
AIC_COMP	44	42	41	36	29	18
PVAL	34	35	31	31	25	18
AIC	34	35	31	31	25	18
L7	30	32	28	29	26	18
L1	30	33	30	27	22	15
LAICO	29	30	26	27	23	19
INTERP	32	31	30	26	18	12
WRTDS	27	27	25	25	24	20
RATIO_F5	27	27	27	25	22	18
RATIO_T	30	29	29	23	18	12
RATIO_F1	33	23	21	25	19	12
L5	20	24	21	21	19	17

Appendix 3. Plots Showing the Distribution of Errors of Annual Load-Estimation Methods among Sampling Strategies

The figures in appendix 3 are available for download at https://doi.org/10.3133/sir20195084.

Figure 3.1. Comparison of INTERP method errors among sampling strategies.

Figure 3.2. Comparison of RATIO_T method errors among sampling strategies.

Figure 3.3. Comparison of RATIO_F1 method errors among sampling strategies.

Figure 3.4. Comparison of RATIO_F5 method errors among sampling strategies.

Figure 3.5. Comparison of L1 method errors among sampling strategies.

Figure 3.6. Comparison of L5 method errors among sampling strategies.

Figure 3.7. Comparison of L7 method errors among sampling strategies.

Figure 3.8. Comparison of LAICO method errors among sampling strategies.

Figure 3.9. Comparison of AIC method errors among sampling strategies.

Figure 3.10. Comparison of PVAL method errors among sampling strategies.

Figure 3.11. Comparison of AIC_COMP method errors among sampling strategies.

Figure 3.12. Comparison of WRTDS method errors among sampling strategies.

Figure 3.13. Comparison of WRTDS_K method errors among sampling strategies.

Appendix 4. Plots Showing the Distribution of Errors of Annual Load-Estimation Methods among Sampling Sites

The figures in appendix 4 are available for download at https://doi.org/10.3133/sir20195084.

Figure 4.1. Comparison of estimation method errors for computing annual chloride loads.

Figure 4.2. Comparison of estimation method errors for computing annual total nitrogen loads.

Figure 4.3. Comparison of estimation method errors for computing annual nitrate plus nitrite loads.

Figure 4.4. Comparison of estimation method errors for computing annual total phosphorus loads.

Figure 4.5. Comparison of estimation method errors for computing annual suspended-sediment loads.

Appendix 5. Evaluation of Estimation Method Performance among Sampling Windows

Analyses in appendixes 5 and 6 are completed to evaluate aspects of load estimation specific to National Water Quality Network load-estimation procedures. An evaluation of the performance of selected methods among different sampling windows to characterize the degree to which historical water-quality observations should be used to estimate loads for a given year is in this appendix. Several methods described previously are not considered in this evaluation because they are not amenable to considering a variety of sampling window lengths. The interpolation of sampled values method and the Beale's ratio estimation with time-based stratification method are not evaluated because they only use data from the target water year by design (a water year is the period from October 1 to September 30 and is designated by the year in which it ends). The Weighted Regressions on Time, Discharge, and Season method and the Weighted Regressions on Time, Discharge, and Season method with Kalman filtering (WRTDS_K) consider a user-specified number of samples by design, and thus are not evaluated among sampling windows. In addition, the LOADEST minimum probability-value model with cubic streamflow and streamflow anomaly variables method and the LOADEST Akaike information criteria method with local adjustments for residual departures using the composite method are not evaluated because results were similar to the LOADEST minimum Akaike information criteria method with additional explanatory variables (AIC) method with respect to sampling windows.

Sampling windows are evaluated for the LOADEST stock 5-parameter model with streamflow, season, and time as explanatory variables (L5); LOADEST stock 7-parameter model with streamflow, streamflow squared, season, time, and time squared as explanatory variables (L7); AIC; and Beale's ratio estimation with streamflow-based stratification (RATIO_F) methods using (1) streamflow and water-quality data for the target water year only, (2) streamflow and water-quality data from the target year and the previous 2 years (a 3-year window), (3) data from the target year and the previous 4 years (a 5-year window), and (4) data from the target year and the previous 6 years (a 7-year window). RATIO_F is the only method that uses data from the target water year only (that is, a 1-year window) as well from 3-, 5-, and 7-year windows (fig. 5.1). The LOADEST stock 1-parameter model with streamflow as the only explanatory variable (L1) method is used to compute 1-year sampling window estimates in figures 5.2, 5.3, and 5.4 because this was the only LOADEST-based method that used a 1-year sampling window. This analysis is completed using only a single replicate for a given site, constituent, and water year because of limitations on processing time. The bimonthly sampling strategy is excluded from this evaluation because the relatively few observations obtained using this strategy occasionally caused LOADEST software to fail.

The figures in appendix 5 are available for download at https://doi.org/10.3133/sir20195084.

Figure 5.1. Comparison of RATIO_F estimation method errors across sampling windows.

Figure 5.2. Comparison of L5 estimation method errors across sampling windows.

Figure 5.3. Comparison of L7 estimation method errors across sampling windows.

Figure 5.4. Comparison of AIC estimation method errors across sampling windows.

Among water-quality constituents, methods, and sampling strategies considered, 1-year sampling windows produced the fewest estimates (59 percent) within plus or minus (±) 20 percent of observed loads. The 3-year (63 percent), 5-year (63 percent), and 7-year (62 percent) windows produced more estimates within this threshold among all methods; however, the degree to which the consideration of data beyond the target year improved estimates varied among estimation methods and water-quality constituents. RATIO_F total nitrogen (80 percent within ±20 percent of observed loads) and nitrate plus nitrite (74 percent) estimates were most accurate using a 1-year window; however, 5- or 7-year windows produced the most within this threshold for chloride, total phosphorus, and suspended sediment. The 1-year window (that is, the L1 method) produced more total nitrogen and suspended-sediment estimates within ±20 percent of observed loads than 3-, 5-, and 7-year sampling window estimates obtained from the L5 method, but nitrate plus nitrite estimates were similar among sampling windows. The 3-, 5-, or 7-year windows produced slightly more accurate chloride and total phosphorus estimates within the ±20-percent threshold (fig. 5.2). The L7 method was most accurate using 3-, 5-, or 7-year sampling windows for chloride, total nitrogen, and nitrate plus nitrite loads; however, the L1 and L7 methods produced similar percentages of estimates within ±20 percent of observed loads for total phosphorus and suspended sediment (fig. 5.3). AIC estimates using 3-, 5-, and 7-year windows generally improved upon L1 estimates for all water-quality constituents. The most extreme deviations from observed values, characterized as estimates greater than 100 percent or −50 percent from observed loads, were disproportionately observed for 1-year sampling windows (8.9 percent of estimates) as compared to 3-, 5-, or 7-year sampling windows among all methods and constituents (6.4 percent, 6.1 percent, and 6.4 percent of estimates, respectively).

Although slight differences were observed, the percentage of estimates within ±20 percent of observed values was similar among 3-, 5-, and 7-year sampling windows among estimation methods and constituents (figs. 5.1, 5.2, 5.3, and 5.4). Generally, comparisons indicate that including water-quality observations beyond the target year improves the accuracy of load estimates. Major exceptions to this finding were the total nitrogen and nitrate plus nitrite estimates, in which ratio estimators computed among the most accurate loads using sample data from the target year only. Based on these results, National Water Quality Network load-estimation procedures will continue to use samples from a 5-year window to compute loads for all water-quality constituents.

Appendix 6. Evaluating Potential Improvements in Method Performance through Graphical Examination of Residuals

As described in Lee and others (2017), the U.S. Geological Survey National Water Quality Network (NWQN) uses an adapted-LOADEST procedure in which the LOADEST minimum Akaike information criteria method with additional explanatory variables (AIC) and the LOADEST minimum probability-value method with additional explanatory variables (PVAL) produce candidate models that are then evaluated by an analyst to select a model that best conforms to regression model assumptions. For a given water year (a water year is the period from October 1 to September 30 and is designated by the year in which it ends), AIC, PVAL, and the model form used in the previous water year (for the same site and water-quality constituent) are evaluated using graphics illustrated in figures 6.1 and 6.2. A total of eight plots, which are used to evaluate model fit in logarithmic and arithmetic space (Lee and others [2017]), are shown in figure 6.1 (adapted from Hirsch and others [2010]). Sampled values relative to model estimates in logarithmic space are shown in figure 6.2; samples are color coded by the water year in which samples were collected to allow the analyst to characterize if observations from the current water year are consistent with those in previous water years. If none of the three candidate models reasonably meet regression assumptions in the view of the analyst, other model forms are plotted and evaluated. The NWQN sampling procedure uses a 5-year moving window as described in appendix 5 and by Lee and others (2017).

Figure 6.1. Eight-panel figure adapted from Hirsch and others (2010) to evaluate models for the National Water Quality Network method (https://doi.org/10.3133/sir20195084).

Figure 6.2. Comparison of observed versus estimated daily constituent loads by water year (https://doi.org/10.3133/sir20195084).

The NWQN method is evaluated in this appendix to determine the degree to which the inspection of model residuals improved (or reduced) the accuracy of loads computed using the AIC and PVAL methods. As with results presented in appendix 5, the bimonthly sampling frequency is excluded from this analysis, and this evaluation is completed using only a single replicate for a given site, constituent, and water year because of limitations on processing time. Although loads are computed at NWQN sites in practice only when an adequate regression model can be identified, loads are computed in all cases in this study for comparative purposes. In cases where model residuals seem similar among methods, models that are relatively unbiased for samples collected at the highest streamflow and loading conditions in the target water year are favored. Decisions regarding which model to use in this NWQN method were made without prior knowledge of how results compared to observed loads.

Inspection of model residuals by the NWQN load-estimation method offered relatively little improvement relative to the AIC and PVAL methods. Among all constituents, the NWQN method produced 71 percent of estimates within plus or minus (±) 20 percent of observed loads; the AIC (70 percent) and PVAL (69 percent) methods produced similar results with respect to this threshold. One potential benefit of the NWQN method is that the inspection of residuals affords the opportunity to identify extremely biased estimates. The most extreme deviations, characterized as estimates greater than 100 percent or less than −50 percent from observed loads, occurred less frequently when using the NWQN method compared to the AIC and PVAL methods. NWQN estimates resulted in extreme errors for 3.5 percent of estimates, whereas the AIC and PVAL methods produced extreme errors in 5.1 percent and 5.3 percent of cases, respectively (fig. 6.3).

Figure 6.3. Comparison of NWQN, AIC, and PVAL method errors among water-quality constituents (https://doi.org/10.3133/sir20195084).

The accuracy of load estimates among the NWQN, AIC, and PVAL methods varied slightly among specific water-quality constituents. The AIC method produced slightly more estimates within ±20 percent of observed loads than the NWQN or PVAL methods for chloride and total nitrogen estimates (fig. 6.3), whereas the NWQN method produced slightly more estimates within this threshold for nitrate plus nitrite, total phosphorus, and suspended-sediment estimates. However, for each constituent, methods generally had similar performance and all three demonstrated the potential to produce accurate or inaccurate water-quality loads. The relative lack of improvement from the examination of model residuals emphasizes that (1) observations only provide a representation of observed conditions for a given target year; (2) a prescribed model form may not be able to adequately characterize relations among water-quality concentrations, streamflow, and time; and (3) the adjustment of daily estimates based on sampled values through the LOADEST Akaike information criteria method with local adjustments for residual departures using the composite method and the Weighted Regressions on Time, Discharge, and Season method with Kalman filtering (WRTDS_K), described previously, offered more potential for improving accuracy as compared to examining model performance graphically. As described in the main text, although the NWQN method will continue to be used to estimate water-quality loads for the purposes of maintaining a consistent historical record, results also will be computed using the WRTDS_K method because it produced the most accurate results among sites, constituents, and water years.

References Cited

Hirsch, R.M., Moyer, D.L., and Archfield, S.A., 2010, Weighted Regressions on Time, Discharge, and Season (WRTDS), with an application to Chesapeake Bay River inputs: Journal of the American Water Resources Association, v. 46, no. 5, p. 857–880. [Also available at https://doi.org/10.1111/j.1752-1688.2010.00482.x.]

Lee, C.J., Murphy, J.C., Crawford, C.G., and Deacon, J.R., 2017, Methods for computing water-quality loads at sites in the U.S. Geological Survey National Water Quality Network: U.S. Geological Survey Open-File Report 2017–1120, 20 p., accessed July 2018 at https://doi.org/10.3133/ofr20171120.

Appendix 7. Description of Methods and Results from Regression-Tree Analyses

This section describes the methods and results of regression-tree analyses used to characterize if sampling record characteristics could predict the accuracy of water-quality load-estimation methods, as summarized in the "Causes of Error among Estimation Methods" section in the main text. A total of 91 aspects of sampling records were computed to evaluate characteristics that may affect the accuracy of annual load estimates for selected methods. Each of these variables could be computed using daily streamflow and (or) periodically collected discrete water-quality data. Aspects of sampling records considered include measures of the variability of sampled concentrations; streamflows; and loads for the target year, for specific streamflow conditions within the target year, and for samples collected over the target year and previous 4 years. Also considered were the percentage differences among mean sampled and daily streamflows for the given water year (for the entire year and for subsets of streamflow conditions; a water year is the period from October 1 to September 30 and is designated by the year in which it ends), the ratio among peak sampled and observed streamflow conditions, the length of the previous sampling record, the slope relations among sampled concentrations and streamflow conditions in logarithmic space (for the entire sampled record and for subsets of streamflow conditions), the coefficient of determination values of linear and quadratic regressions of the logarithm of concentration and streamflows (for the entire sampled record and subsets of streamflow conditions), and the base-flow index (computed using the R EcoHydrology package) for each year at a given sampling site. Because the purpose of this analysis is to distinguish relatively unique aspects of sampling records that affect load-estimate accuracy, and because many of the initial 91 variables considered were correlated among each other, a set of 17 variables representative of different types of record characteristics that were relatively uncorrelated (Pearson correlation coefficients less than 0.8) were selected for further analysis (table 7.1). These variables include the base-flow index; the coefficient of variability of sampled concentrations, loads, and streamflows for the target water year; measures of the representativeness of sampled versus observed streamflows; the number of years of previously sampled records; and the correlation and slopes of linear relations among the logarithm of sampled concentration and streamflows over the most recent 5 years.

Regression-tree analysis was completed to illustrate the potential for sampling record characteristics to predict the accuracy of Weighted Regressions on Time, Discharge, and Season method with Kalman filtering (WRTDS_K), LOADEST minimum Akaike information criteria method with additional explanatory variables and adjustment via the composite method (AIC_COMP), and LOADEST minimum Akaike information criteria method with additional explanatory variables (AIC) load estimates. AIC estimates are considered in addition to the WRTDS_K and AIC_COMP methods (which were generally the most accurate) to characterize how factors affecting load-estimate accuracy may differ among methods that do and do not use localized adjustments based on departures from sampled values. Although regression trees can provide useful visualizations of how factors affect a dependent variable, results among subsets of a given dataset can be highly variable and affected by outlying observations. Several steps were taken to address these limitations in this analysis. First, regression trees were computed from training datasets that consist of samples that were randomly selected from 90 percent of the original dataset to assess how trees varied among training datasets and to allow the accuracy of trees to be quantified using the remaining data. Second, to discount the effect of outlying data, load-estimate accuracy is characterized using a categorical threshold. Thresholds are allowed to vary among estimation methods and water-quality constituents such that about 50 percent of estimates closest to observed loads are defined as "accurate," whereas the remaining estimates are defined as "inaccurate." This approach enables results to be more readily compared among methods and constituents. Depending on the method and constituent, the accuracy threshold ranged from plus or minus (±) 6 to 16 percent within observed loads. Third, the average overall model accuracy and the importance of explanatory variables are assessed using a bootstrap aggregating process (termed "bagging") in which regression-tree results are averaged across 50 replicates computed using random samples extracted from 90 percent of the training dataset. The importance (with 1 being the most important) of explanatory variables among estimation methods and water-quality constituents is ranked in table 7.2. Example regression trees that are split no more than three times are shown in fig. 7.1 to provide a simplified illustration of how explanatory variables typically interact to predict load-estimate accuracy. It is important to note that because of interactions among explanatory variables and limits set on tree length, variables shown in fig. 7.1 will not necessarily reflect the most influential variables shown in table 7.2. Regression-tree analysis is completed using high-flow sampling and biweekly sampling estimates only; trees are computed from each method using estimates from all constituents (chloride, total nitrogen, nitrate plus nitrite, total phosphorus, and suspended sediment) and individually for chloride, total phosphorus, and suspended sediment.

Among all constituents, WRTDS_K estimates were approximately evenly divided within or outside of ±8 percent of observed loads (fig. 7.1). Among 50 bootstrapped estimates, average "out-of-bag" regression-tree predictions (that is, the average of those not in bootstrapped samples) correctly placed 68 percent of estimates as within or outside of the ±8-percent threshold. The variability of sampled concentrations (CV_C), the variability of sampled concentrations at the highest

Table 7.1. Variables selected for regression-tree analysis.

[R^2, coefficient of determination]

Variable	Definition
BFI	Base-flow index of the observed flow record for a given site and water year.
CV_C	Coefficient of variation of sampled concentrations for a given site, constituent, and water year.
*CV_C*30	Coefficient of variation of the top 30 percent sampled concentrations collected at the highest flow conditions for a given site, constituent, and water year.
CV_L	Coefficient of variation of sampled loads for a given site, constituent, and water year.
CV_F	Coefficient of variation of sampled flows for a given site, constituent, and water year.
MEAN_FLOW	Mean percentage difference of sampled flows from observed sampled flows for a given site, constituent water year.
*MEAN_FLOW_B*50	Mean percentage difference of sampled flows from observed sampled flows for the bottom 50 percent of flows for a given site, constituent, and water year.
*MEAN_FLOW_B*10	Mean percentage difference of sampled flows from observed sampled flows for the bottom 10 percent of flows for a given site, constituent, and water year.
NYRS	Number of previously sampled years for a given site, constituent, and water year.
FLOW_PK	Percentage difference between peak sampled and peak observed flow for a given site, constituent, and water year.
*R*2	R^2 among the log of sampled concentration and flows for the most recent 5 water years for a given site and constituent.
*R*2_10	R^2 among the log of sampled concentration and flows for the 10 percent of samples at the highest flows for the most recent 5 water years for a given site and constituent.
*R*2_50	R^2 among the log of sampled concentration and flows for the 50 percent of samples at the highest flows for the most recent 5 water years for a given site and constituent.
*R*2_*DIFF*50	Absolute value of the difference between the *R*2 and *R*2_50 variables.
SL	Slope of the log of sampled concentration and sampled flows for the most recent 5 water years for a given site and constituent.
*SL*10	Slope of the log of sampled concentration and flows for the 10 percent of samples at the highest flows for the most recent 5 water years for a given site and constituent.
*SL*50	Slope of the log of sampled concentration and flows for the 50 percent of samples at the highest flows for the most recent 5 water years for a given site and constituent.
*SL_DIFF*50	Absolute value of the difference between the *SL* and *SL*50 variables.

Table 7.2. Ranked importance of explanatory variables in predicting load-estimate accuracy [variables are ranked from most important (1) to least important (18)].

[WRDTS_K, Weighted Regressions on Time, Discharge, and Season method with Kalman filtering; AIC_COMP, Akaike information criteria method with an additional adjustment of daily estimates by the composite method (Aulenbach and Hooper, 2006); AIC, LOADEST minimum Akaike information criteria method with additional explanatory variables as described in Lee and others (2017)]

Variable	WRTDS_K				AIC_COMP				AIC			
	All	Nitrate plus nitrite	Total phosphorus	Suspended sediment	All	Nitrate plus nitrite	Total phosphorus	Suspended sediment	All	Nitrate plus nitrite	Total phosphorus	Suspended sediment
BFI	14	2	6	1	12	3	2	1	10	1	1	1
CV_C	1	6	1	3	1	4	1	3	1	4	2	3
*CV_C*30	2	7	3	6	5	6	5	5	8	7	9	12
CV_L	6	4	8	4	2	8	7	4	4	15	4	6
CV_F	5	5	7	8	7	5	6	7	5	11	5	4
MEAN_FLOW	9	3	5	11	8	2	4	11	11	12	12	13
*MEAN*_FLOW_B50	11	11	16	15	11	10	11	12	13	17	10	16
*MEAN_FLOW_B*10	13	15	15	16	13	17	15	15	14	18	18	18
NYRS	17	18	18	18	17	18	18	18	12	5	11	8
FLOW_PK	8	1	2	2	6	1	3	2	7	10	7	2
*R*2	12	16	10	5	10	13	13	6	9	6	8	5
*R*2_10	18	8	17	17	18	15	14	17	18	13	16	17
*R*2_50	16	12	11	10	16	16	8	10	15	14	15	11
*R2_DIFF*50	15	14	13	13	14	14	17	13	16	9	14	10
SL	4	10	9	9	3	9	9	9	2	2	6	7
*SL*10	7	9	14	14	15	12	12	16	17	16	17	14
*SL*50	3	13	4	7	4	11	10	8	3	8	3	9
*SL_DIFF*50	10	17	12	12	9	7	16	14	6	3	13	15

30 percent of streamflows within the sampled record (*CV*_30), the slope of concentration/streamflow relations among the 50 percent of target-year samples collected at the highest streamflows (*SL*50), and the slope of concentration/streamflow relations (*SL*) were the most important predictors of whether or not estimates were within ±8 percent of observed loads (table 7.2). Interactions among variables resulted in some different variables being used in practice to best characterize WRTDS_K estimate accuracy. Individual regression trees for WRTDS_K were typically first divided based on increasing variability in sampled concentrations (*CV_C*); sites, constituents, and years with *CV_C* values less than 76 produced loads within ±8 percent of observed loads in 64 percent of cases, whereas those more variable concentrations produced loads within the ±8-percent threshold for 37 percent of cases (fig. 7.1). Among records with less variable sampled concentrations, those with higher ratios of peak sampled/observed streamflows (that is, better representation of peak-flow conditions) were within ±8 percent of observed loads more often than records with smaller ratios of peak sampled/observed streamflows. Among cases with less variable concentrations and less representative sampling of peak streamflow conditions, lower slopes among concentration/streamflow relations at higher streamflows (*SL*50) were within ±8 percent of observed loads more often than those with higher slopes (which is more typical of total phosphorus and suspended-sediment estimates).

Figure 7.1. Example regression trees illustrating relations among estimation method accuracy and explanatory variables (available for download at https://doi.org/10.3133/sir20195084).

As with WRTDS_K estimates, a threshold of ±8 percent approximately evenly divided AIC_COMP estimates into "accurate" and "inaccurate" categories among all constituents. Average out-of-bag regression-tree predictions of the 50 bootstrapped estimates correctly categorized 71 percent of AIC_COMP estimates. The *CV_C*, mean daily streamflow conditions, and the slope of log-log concentration/streamflow relations, for *SL* and *SL*50 were the most important predictors of AIC_COMP estimate accuracy. Typical regression trees (fig. 7.2) were first divided by *CV_C*; sites/constituents/years with *CV_C* values less than 60 were within ±8 percent of observed loads 71 percent of the time, whereas those higher slopes were within this threshold only 37 percent of the time. Among sites/constituents/years with higher *CV_C* values, those with sampled streamflows that more closely approximated peak observed streamflows (*FLOW_PK*) tended to be more accurate than those with smaller streamflow_*PK* values. Among records with more variable concentrations and higher streamflow_*PK* values, the variability of sampled concentrations again distinguished records; those with more variable concentrations were within ±8 percent of observed loads 36 percent of the time, whereas those with less variable concentrations were within the threshold 54 percent of the time.

AIC-computed loads were approximately evenly divided among those within or outside of ±12 percent of observed loads; out-of-bag regression-tree estimates correctly categorized 72 percent of loads using this threshold. The most important variables were *CV_C*, *SL*, *SL*50, and the variation of sampled loads (*CV_L*). Regression trees were typically broken first by *SL*; higher sloped records were within ±12 percent of observed loads for 41 percent of estimates; records with smaller slopes produced "good" loads for 79 percent of estimates. Among higher sloped records, those with more variability in sampled concentrations (*CV_C*) were typically less accurate than sites/constituents/years with smaller *CV_C* values. Among higher sloped, less variable sampling records, those with more base streamflow (higher *BFI*) were within ±12 percent of observed loads for 54 percent of cases, whereas those with less base streamflow were within this threshold 37 percent of the time.

Although the regression-tree analysis only correctly categorized 68–72 percent of load estimates above the initial approximate 50-percent split of "accurate" and "inaccurate" loads, regression trees were relatively consistent regarding which sampling record characteristics predicted accurate load estimates. More variable concentrations and loads, more runoff (that is, smaller *BFI*s), higher slopes among concentration and streamflow values, and less representation of peak-flow conditions generally led to less accurate load estimates. More variable water-quality concentrations and higher sloped concentration/streamflow relations were typically related to relatively inaccurate water-quality load estimates when considering all constituents. This finding corresponds to previously shown results (table 2) in which most methods accurately computed chloride loads, which tend to have negative slopes and less variable concentrations, relative to total phosphorus and suspended-sediment loads, which tend to have higher slopes and more variable concentrations. When considering nitrate plus nitrite estimates exclusively, more runoff (as illustrated by smaller *BFI*s) and the degree to which sampling represented streamflow_*PK* were the best predictors of relatively accurate/inaccurate load estimates. The variability in sampled concentrations, *BFI*, and streamflow_*PK* were generally the best predictors of increased bias in total phosphorus and suspended-sediment loads. As indicated previously, increased runoff and more variable concentrations result in fewer days transporting most annual water-quality loads. Improved sampling of these peak-flow conditions (that is, higher streamflow_*PK* values) tended to improve the likelihood of producing relatively accurate load estimates for these constituents. Although regression-tree analyses offered some insights regarding factors that contribute to computing biased load estimates, correctly categorizing 68–72 percent of load estimates only represents an approximate 36–44-percent improvement over the initial 50-percent split of "accurate" and "inaccurate" loads. The consideration of alternate sampling record characteristics and (or) use of different techniques may offer an improved ability to identify biased estimates.

References Cited

Aulenbach, B.T., and Hooper, R.P., 2006, The composite method—An improved method for stream-water solute load estimation: Hydrological Processes, v. 20, no. 14, p. 3029–3047. [Also available at https://doi.org/10.1002/hyp.6147.]

Lee, C.J., Murphy, J.C., Crawford, C.G., and Deacon, J.R., 2017, Methods for computing water-quality loads at sites in the U.S. Geological Survey National Water Quality Network: U.S. Geological Survey Open-File Report 2017–1120, 20 p., accessed July 2018 at https://doi.org/10.3133/ofr20171120.

Publishing support provided by:
Rolla Publishing Service Center

For more information concerning this report, contact:
Chief, National Water-Quality Assessment Program
U.S. Geological Survey
413 National Center
12201 Sunrise Valley Drive
Reston, VA 20192
https://water.usgs.gov/nawqa/

ISSN 2328-0328 (online)
https://doi.org/10.3133/sir20195084