Unbiased Estimates and Confidence Intervals for Riverine Loads

Tada, Akio

Tanakamaru, Haruya

(Citation)
Water Resources Research,57(3):e2020WR028170

(Issue Date)
2021-03

(Resource Type)
journal article

(Version)
Version of Record

AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2020WR028170

Unbiased Estimates and Confidence Intervals for Riverine Loads

Akio Tada[1] and **Haruya Tanakamaru**[1]

[1]Graduate School of Agricultural Science, Kobe University, Kobe, Hyogo, Japan

Key Points:

- The unbiased riverine load estimator is theoretically explained based on the rating curve method using importance sampling
- The effectiveness of unbiased point estimation and interval estimation is demonstrated for river loads from various watersheds
- Difficulties in implementing this method at monitoring sites are discussed

Supporting Information:

- Supporting Information S1

Correspondence to:
A. Tada,
atada@kobe-u.ac.jp

Citation:
Tada, A., & Tanakamaru, H. (2021). Unbiased estimates and confidence intervals for riverine loads. *Water Resources Research*, *57*, e2020WR028170. https://doi.org/10.1029/2020WR028170

Received 17 JUN 2020
Accepted 14 FEB 2021

Abstract Estimating the uncertainty in annual riverine constituent loads, which is the mass that passes through a river cross-section into a receiving water body, using infrequent water quality (WQ) observations is a difficult and unsolved task. Therefore, we propose an unbiased point estimation method and interval estimation method for river loads based on the rating curve (RC) method using importance sampling and the bootstrap method, respectively. In this paper, we first statistically explain the unbiasedness of load estimates using the proposed method. Second, the effectiveness of point and interval estimates by the proposed method is demonstrated for river loads from a small catchment and from large watersheds based on discharge and WQ data of solutes, nutrients, and suspended sediments with 10-min to daily intervals. The results show that the proposed method provides unbiased estimates and appropriate coverage of confidence intervals regardless of the RC model used. The results also reveal that the dominant cause of bias in load estimates based on ordinary RC methods, such as the Loadest model, is not due to the poor simulation of observed loading rates by the RCs or because of the nonnormality of the regression residuals but rather improper sampling strategies. The proposed method is currently not feasible for WQ monitoring sites in large rivers due to the unmanageability of missing observations or censored data and an inefficient sampling strategy, although the requirement of unbiased estimation explained here can aid in scheduling high-flow sampling for monitoring sites.

1. Introduction

An accurate estimation method for seasonal, annual, or decadal riverine constituent loads, which are the masses that pass through a river cross-section into a receiving water body during target periods, has been investigated on the premise of high-frequency discharge and infrequent water quality (WQ) observations (e.g., Philips et al., 1999; Webb et al., 1997) due to the high cost and labor of WQ monitoring and analysis (Horowitz, 2013; Verma et al., 2012). Load estimation studies have targeted the WQ parameters related to water pollution or erosion, such as nutrients or suspended sediments (SS), from catchments or watersheds with various land uses. Among the load estimation methods (LEMs), the rating curve (RC) method (RCM) is very common, whereas the river load estimates of nutrients and SS by this method can contain large biases and variances (e.g., Lee et al., 2016; Walling & Webb, 1981). Load estimation studies have also retrospectively evaluated the uncertainty in the load estimates by several LEMs, including the RCM using the Monte Carlo method based on high-frequency WQ data sets, daily data sets (Johnes, 2007; Lee et al., 2016, 2019; Preston et al., 1989) or subhourly to daily data sets (Birgand et al., 2010). Many such studies, which have attempted to empirically evaluate the bias of the load estimates by the tested LEMs, have aimed at obtaining reliable estimates with the specified uncertainty using small-size WQ samples (Gulati et al., 2014; Lessels & Bishop, 2020). However, their findings are summarized as follows: it is difficult to construct a generalized and universal LEM that can provide accurate load estimates (Hirsch, 2014; Lee et al., 2016; Schleppi et al., 2006) because load estimates are subject to the employed load calculation method (LCM), the sample sizes or sampling frequency of the monitoring scheme (e.g., Philips et al., 1999), the watershed size (e.g., Birgand et al., 2010), and the behavior of WQ parameters (Verma et al., 2012).

In the history of load estimation using RCs, log-transformation bias (Cohn et al., 1989; Ferguson, 1986), nonnormality of residuals by the RCs (Koch & Smillie, 1986), and the heteroscedasticity of residuals (Hirsch, 2014; Stenback et al., 2011) have been considered as the major causes of bias in load estimates by the traditional RCM. In other words, the cause of the bias in load estimates by the RCM has been attributed to the violation of traditional assumptions of the linear regression model: normality of errors with a

zero mean, no systematic errors by the RCs, homoscedasticity of errors, and the independence of errors. Although an unbiased LEM that can be utilized with an RCM has been developed, named the selection at list time (SALT) method, based on probability sampling (Thomas, 1985), this method has never been employed at WQ monitoring sites due to the requirement of a sophisticated autosampler that enables random sampling at monitoring sites (Cohn et al., 1992). This LEM was based on sampling with a probability that is proportional to size (PPS, Hansen & Hurwitz, 1943; Raj, 1954). Even in Thomas' study, a theoretical explanation of unbiased load estimates and validation based on highly frequent WQ observations have not been provided.

In addition to the bias of load estimates, the reliability or uncertainty of each load estimate is important in studies about water body protection from pollution or watershed-scale mass balance. Because we cannot know the deviation between each estimate and the true load value, information on the magnitude of the uncertainty in each estimate is essential in such studies. Although a confidence interval (CI) can provide such information on the uncertainty in an estimate, past studies that evaluated CIs of load estimates have made incomplete assumptions that load estimates would follow a lognormal distribution (Cohn, 2005) or a normal distribution; normality of total loads (Cohn, 2005; Thomas, 1985); or normality of average daily loads (International Reference Group on Great Lakes Pollution from Land Use Activities & Whitt, 1977; Verhoff et al., 1980). Weak points in these assumptions made for RCMs were discussed in Appling et al. (2015): the assumption of a normal distribution of load estimates by the central limit theorem requires large samples. The assumption of a lognormal distribution of load estimates also conflicts with the normality assumption for a distribution of regression residuals around the RC in log space or with a lognormality assumption for a distribution of loading rates. Other techniques, namely, the bootstrap method, such as bootstrap percentile CIs (Mailhot et al., 2008; Rustomji & Wilkinson, 2008; Slates et al., 2017) and bias-corrected and accelerated bootstrap (BCa) CIs (Vigiak & Bende-Michl, 2013), or Bayesian estimation (Pagendam et al., 2014; Vigiak & Bende-Michl, 2013), have been used to construct CIs, prediction intervals, or credible intervals of load estimates. Vigiak and Bende-Michl (2013) reported coverage of intervals, which are the probabilities of intervals bracketing the reference load value or the true load value, of the 95% bootstrap prediction intervals and 95% Bayesian credible intervals for daily and monthly load estimates, but their reported coverage was not comparable with the 0.95 for all WQ parameters evaluated. We have not yet established a method for evaluating the uncertainty in load estimates.

The unsatisfactory performance of those reported intervals and failure to find a universally less biased LEM arise from the use of a biased LEM or insufficient understanding of bias itself. Even in many studies about accurate load estimation, few studies (Koch & Smillie, 1986; Thomas, 1985) have discussed unbiased estimation according to the statistical definition. The word "unbiased" or "accurate" was often misused as "likely or probably less biased" even though the word "bias" was properly defined as the difference between the expected value of estimates and the true value. To avoid such confusion, we need to define the unbiased load estimator here. According to statistical textbooks, such as Keener (2010), the unbiased load estimator refers to the one whose expected value of estimates agrees with the true value L_{TRUE} regardless of the model parameters or regression coefficients θ used for the estimation. That is,

$$E_\theta\left[\hat{L}\right] = L_{\mathrm{TRUE}}, \quad \forall\theta \tag{1}$$

where $\hat{L}$ is a load estimate and $E_\theta[\cdot]$ denotes the expectation of $\cdot$ under model parameter θ. The importance of Equation 1 comes from the fact that we can determine neither the best model structure for the population nor the model parameters for the population based on small samples. That is, the expectation of the estimate and the true value need to be matched even when the biased model parameters are used for estimation, as in Equation 1. Thomas (1985) suggested that the approximation quality of the loading rates, that is, the deviation of the model parameter values from the population parameters, did not affect the unbiasedness (lack of bias) of the load estimator, and it did affect the variance in the load estimates by the SALT method. This description ensures Thomas' proper understanding regarding the unbiased load estimator. To properly estimate the uncertainty in load estimates, CIs must be based on the unbiased estimator because a biased estimator lowers the coverage of CIs relative to the expected significance level.

Based on the above review, this study proposes a configuration method for the statistically appropriate CIs of load estimates based on an unbiased LEM using the RCM, which is most widely used in load estimation, and its effectiveness is evaluated using the observed high-frequency WQ monitoring data for various parameters from various watersheds. For these purposes, we first statistically explain and propose the unbiased LEM using the RCM based on importance sampling (IS), which is an efficient Monte Carlo numerical integration method, and we demonstrate that the proposed unbiased LEM (RCM using IS) is identical to the LEM based on the SALT method. Then, we propose the CI configuration method for a river load estimate based on the RCM using IS and show the effectiveness of the proposed interval estimation method and unbiasedness of load estimates for river loads from a small catchment in Japan and from large watersheds in the United States based on discharge and WQ data of solutes, nutrients, and SS with 10-min to daily intervals. We also discuss the cause of bias in load estimates by the traditional RCM by comparing them with estimates by the unbiased LEM. Finally, in addition to the effectiveness of the LEM, the feasibility of the proposed LEM at WQ monitoring sites is examined, as the application of the LEM to monitoring sites requires solutions to problems related to sample collection and incomplete observations.

Additionally, we must define the range of the uncertainty discussed in this paper: we focus on load estimation uncertainties derived from the sampling strategy and the LCM. This paper did not focus on uncertainties resulting from discharge observation errors (errors in a stage-discharge RC or water stage measurement; Hollaway et al., 2018; Lloyd et al., 2016), errors associated with the WQ measurement (McMillan et al., 2012), representativeness of sampling points in the river cross-section (Horowitz, 2013; McMillan et al., 2012; Rode & Suhr, 2007), or integrated uncertainty that includes all of the above factors (Harmel et al, 2006, 2009; Horowitz, 2013; Yanai et al., 2015). However, the effects of missing observations and censored data that contain observations below the detection limit, on load estimation by the proposed method, are discussed in this paper in connection with the feasibility of the proposed method at WQ monitoring sites.

2. Materials and Methods

2.1. Water Quality and Discharge Data

In this paper, we used four data sets from a small forested headwater catchment in Japan and six data sets from four watersheds affected by anthropogenic activities, such as agriculture. In principle, the evaluation of the effectiveness of an LEM based on bias and the coverage of CIs of load estimates requires use of the reference or true load value for the target period. To calculate the reference or true load, WQ monitoring data with short intervals, at which river WQ concentration and discharge are approximated as constants, are required. To approximately meet this requirement, we used K, Cl, Na, and suspended sediment (SS) data and discharge data with 10-min intervals from a small forested catchment (Gozyo) of 12.14 ha in Nara, Japan (Table 1). The observation periods are May 12, 2009–April 28, 2011 and June 7, 2012–August 26, 2014 for solutes and SS, respectively. Because the uncertainty in the load estimates can be greater as the catchment size decreases (Moatar et al., 2006) due to larger variations in WQ and discharge by rapid hydrological responses (Scleppi et al., 2006), high-frequency WQ data sets from a small catchment can provide a good foundation for testing the proposed LEM.

In the Gozyo catchment, solute concentrations were measured using an on-site flow injection potentiometry (FIP) system (Kurahashi Giken Corp., FIA-A, and FIA-S) with ion selective electrodes at 15 min intervals required to avoid sample carry-over contamination (Tada et al., 2006). Solute data sets with 10-min intervals were interpolated from raw monitoring data. The discharge data were calculated in 10-min intervals from continuous records of the stream stage in the V-shaped weir at the catchment outlet using the stage-discharge RC. Turbidity data were obtained in 10-min intervals at the catchment outlet using an on-site integrating sphere turbidity meter (Tohken Engineering Co., Ltd., RT530-T) from 2012 to 2016. The SS concentrations of the 163 samples collected from 2012 to 2014 by the automatic sampler were also measured to establish the relationship between the SS and turbidity. The linear correlation between them was recognized as $SSC = 1.09 \times T_b + 6.19$, and the coefficient of determination R^2 was 0.909, where SSC and T_b are the SS concentration in mg/L and the turbidity in nephelometric turbidity units, respectively. Using this relationship, SS concentrations were calculated from turbidity data in 10-min intervals. In these data sets, there were some periods of missing data due to battery troubles or temporal system malfunctions for solutes

Table 1
Sites and Water Quality Parameters Used in the Load Evaluation

Site name (USGS gauge number[a])	Abbreviation	WQ parameter	Period of record used[b]	Drainage area (km^2)[c]	Land use (%)[c]			Data interval	Data source	ND (%)[d]	L_{ref}[e]	Period of L_{ref}
					A	F	U					
Muskingum River (03150000)	MK-N	Nitrite & nitrate	WY2008–WY2018	19,223	24	43	12	1 [d]	NCWQR[f]	2.8	107.9	WY2009–WY2018
Rock Creek (04197170)	RC-TP1	Total phosphorus	WY1988–WY1998	90.65	72	11	9	1 [d]	NCWQR[f]	3.8	141.6	WY1989–WY1998
Rock Creek (04197170)	RC-TP2	Total phosphorus	WY1998–WY2008	90.65	72	11	9	1 [d]	NCWQR[f]	5.5	128.1	WY1999–WY2008
Rock Creek (04197170)	RC-TP3	Total phosphorus	WY2008–WY2018	90.65	72	11	9	1 [d]	NCWQR[f]	3.0	154.0	WY2009–WY2018
Skunk River (05474000)	SK-SS	Suspended sediment	WY2008–WY2018	11,168	77	7	7	1 [d]	USGS	0.0	20.015	WY2009–WY2018
Vermilion River (04195000)	VM-N	Nitrite & nitrate	WY2002–WY2007	679	73	25	1	1 [d]	NCWQR[f]	15.0	3.340	WY2003–WY2007
Gozyo (Nara)	GZ-Cl	Chloride	2009–2011	0.1214	0	100	0	10 [min]	Repository[g]	18.9	0.391	2009–2011
Gozyo (Nara)	GZ-K	Potassium	2009–2011	0.1214	0	100	0	10 [min]	Repository[g]	20.5	0.121	2009–2011
Gozyo (Nara)	GZ-Na	Sodium	2009–2011	0.1214	0	100	0	10 [min]	Repository[g]	30.8	0.358	2009–2011
Gozyo (Nara)	GZ-SS	Suspended sediment	2012–2014	0.1214	0	100	0	10 [min]	Repository[g]	96.8	15.001	2012–2014

[a]Site name with USGS gauge number for U.S. watersheds and prefecture name for the catchment in Japan. [b]WY, Water year from October 1 to September 30. [c]Information on US sites is from Heidelberg University National Center for Water Quality Research (https://ncwqr.files.wordpress.com/2020/01/basic-station-information-ao-wy2019.xlsx) and Lee et al. (2016). A, Agriculture; F, Forest; U, Urban. [d]Percent of missing or undetected water quality/discharge data. [e]Units of the reference loads L_{ref} are [kt] for MK-N and VM-N, [t] for RC-TP1/2/3 and Gozyo data, and [Mt] for SK-SS. [f]Heidelberg University National Center for Water Quality Research. [g]Zenodo repository (https://doi.org/10.5281/zenodo.4485600).

and zero turbidity, that is, observations below the detection limit, reported in clear water in low flows for SS. The percentages of nonzero concentrations were 79.5%, 81.1%, 69.2%, and 3.2% for K, Cl, Na, and SS, respectively. We used only nonzero concentration data in the load evaluation. Henceforth, these 10-min K, Cl, Na, and SS data sets are referred to as GZ-K, GZ-Cl, GZ-Na, and GZ-SS, respectively. The Gozyo data sets are available from the Zenodo repository (https://doi.org/10.5281/zenodo.4485600). The reference load L_{ref}, which is the surrogate for the true load L_{TRUE} used in the load evaluation, was the total sum of the loading rates, each of which is defined as the product of observed discharge and nonzero concentration value, in 10-min intervals for the target period.

In addition, we used six data sets from four watersheds with various land uses to test the effectiveness of the proposed LEM (Table 1). Although the Gozyo data sets can provide high-frequency WQ data, the catchment is uniform in its land use and geology. Moreover, load estimation from large watersheds is also considered to be difficult for the following reasons: heterogeneity in land use and geology, anthropogenic activities, such as agriculture, and effects of storage or sweeping of the mass in a river channel. Furthermore, target WQ parameters in water pollution problems are usually nutrients, organic matter, pathogens, pesticides, and SS. For the above reasons, we used the daily discharge and WQ data sets from four agricultural watersheds in the United States.

Daily discharge data sets from four US watersheds and the SS daily data set from the Skunk River watershed, IA (SK-SS), were downloaded from the USGS National Water Information System (NWIS, U.S. Geological Survey, 2016). The other WQ data sets were downloaded from Heidelberg University's National Center for Water Quality Research (NCWQR, Heidelberg University, 2019). We chose the following WQ parameters and watersheds from the NCWQR data sets: nitrite and nitrate data sets from the Muskingum River watershed, OH (MK-N), total phosphorus from the Rock Creek watershed, OH (RC-TP), and nitrite and nitrate from the Vermilion River watershed, OH (VM-N). These data sets were chosen from those with a large bias of load estimates reported in Lee et al. (2016). The lengths of the data sets are 11, 31, 11, and

6-year periods for MK-N, RC-TP, SK-SS, and VM-N, respectively. The RC-TP data sets were divided into three 11-year period data sets, RC-TP1, RC-TP2, and RC-P3, as shown in Table 1, to evaluate the effectiveness of the proposed LEM for a long-term record. The zero or negative value concentrations were excluded from the population data, as in the Gozyo data set. Since the WQ data sets from NCWQR sometimes contain several concentrations in a day, we followed the procedure described in Appendix S1 in Hirsch (2014) to synthesize the daily mean concentration. That is, we calculated daily concentrations as the flow-weighted averages and employed the daily USGS discharge data for the NCWQR data sets. To maintain the replicability of the calculation results, all the data from the four US watersheds are available from the same repository as the Gozyo data sets.

It is noteworthy that the reference load calculated as the sum of daily loading rates can have considerable bias from the true load even though the watershed sizes range from 679 to 19,223 km^2, as shown in Table 1. Because discharge and concentration would show diurnal variation, especially in high flows even in such large watersheds, daily intervals would cause a nonnegligible bias in the calculated reference load relative to the true load. Lee et al. (2017) demonstrated that the reference load values of phosphorous and SS could differ significantly between hourly and daily discharge time units. However, our purpose is not to estimate the true loads for these watersheds but to evaluate the effectiveness of our LEM. For this purpose, we can use daily discharge and WQ data sets in the load evaluation. On the other hand, the exclusion of missing observations or observations below the detection limit from the target period of the load estimation makes it impossible for the RCM using IS to estimate the load for the entire observation period, which is actually needed by the monitoring managers. That is, defining population data by excluding missing or undetected observations from the observation period leads to a feasibility problem at WQ monitoring sites. We will discuss this feasibility problem in Section 4.2.

2.2. Unbiased Load Estimation

2.2.1. Load Estimation Bias by the Ordinary RCM

First, we define the framework of the load estimation problem and then explain the bias of load estimates by traditional RCMs. Suppose that we estimate the load from time t_1 to t_2, and the period $[t_1, t_2]$ can be divided into N constant time intervals (Δts) between observations. We assume that Δt is short so that the river WQ concentration and discharge can be regarded as constant in Δt. Let $\mathbf{Y}_N$ be the vector of N logarithms of observed loading rates l_i ($i = 1$ to N), $\boldsymbol{\beta}_m$ be the vector of M regression coefficients β_j ($j = 0$ to $M-1$) of the RC, and $\mathbf{X}_N$ be the $N \times M$ design matrix whose row vector contains M values of explanatory variables. Subscript m means the values of the population. Let $\mathbf{e}_N$ be the vector of N regression residuals e_i ($i = 1$ to N). For the population of size N, the RC model is expressed as follows:

$$\mathbf{Y}_N = \mathbf{X}_N \boldsymbol{\beta}_m + \mathbf{e}_N \tag{2}$$

Let $\mathbf{x}_i$ be an ith row vector of $\mathbf{X}_N$ and $\mathbf{x}_i$ may consist of the logarithm of discharge or decimal time. For example, $\mathbf{Y}_N = (\ln l_1, \ln l_2, ..., \ln l_N)^{\mathrm{T}}$, $\boldsymbol{\beta}_m = (\beta_{\mathrm{m0}}, \beta_{\mathrm{m1}})^T$, $\mathbf{x}_i = (1, \ln q_i)$, and $\mathbf{e}_N = (e_1, e_2, ..., e_N)^T$, for the two-parameter power law RC, where q is the discharge and superscript T denotes transposition of the vector. In this case, the reference load L_{ref} is calculated as follows:

$$L_{\mathrm{ref}} = \sum_{i=1}^{N} \exp\left(\mathbf{x}_i \boldsymbol{\beta}_m + e_i\right) = \sum_{i=1}^{N} l_i \cong \int_{t_1}^{t_2} l\left(t\right) dt = L_{\mathrm{TRUE}} \tag{3}$$

where $l(t)$ is the observed loading rate at time t.

The calculation of L_{TRUE} or L_{ref} based on Equation 3 requires high-frequency monitoring data, and such WQ monitoring is difficult to carry out due to the high costs of WQ monitoring and analysis. Instead, we usually collect n ($n << N$) samples infrequently and use them in the RCM. Let $\mathbf{Y}_n$ be the vector of n logarithms of sample loading rates and $\mathbf{X}_n$ be the $n \times M$ design matrix. In addition, let $\hat{\boldsymbol{\beta}}$ be the vector of M regressed coefficients $\hat{\beta}_j$ of the RC and $\boldsymbol{\delta}_n$ be the vector of n regression residuals δ_i. Ordinarily, $\hat{\boldsymbol{\beta}}$ is determined using the ordinary least squares (OLS) method based on $\mathbf{Y}_n$ and $\mathbf{X}_n$ under the assumption of a normal distribution of δ_i. For n samples, the RC is expressed as follows:

$$\mathbf{Y}_n = \mathbf{X}_n\hat{\boldsymbol{\beta}} + \boldsymbol{\delta}_n \tag{4}$$

Usually, $\hat{\boldsymbol{\beta}}$ has an estimation error and does not agree with $\boldsymbol{\beta}_m$. Because the logarithm of the ith observed loading rate $\ln l_i$ is expressed as either $\mathbf{x}_i\boldsymbol{\beta}_m + e_i$ or $\mathbf{x}_i\hat{\boldsymbol{\beta}} + \delta_i$, we obtain $\delta_i = \mathbf{x}_i \times (\boldsymbol{\beta}_m - \hat{\boldsymbol{\beta}}) + e_i$. If we do not create any assumption about the distributions of e_i and δ_i, we can estimate the river load L using the non-parametric smearing estimator:

$$\hat{L} = \left[\sum_{i=1}^{N}\exp\left(\mathbf{x}_i\hat{\boldsymbol{\beta}}\right)\right]\left(\frac{1}{n}\sum_{j=1}^{n}\exp\delta_j\right) \tag{5}$$

Although there are other bias correction factors (BCFs) in addition to the smearing estimator, we consider the LEM using Equations 4 and 5 with representative BCFs to be an ordinary RCM in this paper. From Equation 5, the distribution of $\hat{L}$ is subject to the distribution of $(\boldsymbol{\beta}_m - \hat{\boldsymbol{\beta}})$ and e_i. Although Ferguson (1986) and Wang et al. (2011) ignored the effect of the uncertainty in the model parameters, this effect can be substantial, especially when sample size n is small; therefore, we cannot ignore this uncertainty and the effect of the distribution of e_i. The expectation of $\hat{L}$ in Equation 5 becomes the following equation:

$$E\left[\hat{L}\right] = \left[\sum_{i=1}^{N}\exp\left(\mathbf{x}_i\hat{\boldsymbol{\beta}}\right)\right] \times E\left[\exp\left(\mathbf{x}_i\boldsymbol{\beta}_m - \mathbf{x}_i\hat{\boldsymbol{\beta}} + e_i\right)\right] \tag{6}$$

Equation 6 is not identical to Equation 3; therefore, $\hat{L}$ by the ordinary RCM is a biased estimator. Even when $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_m$ and N samples are used, Equation 6 cannot yield the L_{ref} value because $\exp(\mathbf{x}_i\hat{\boldsymbol{\beta}})$ is not independent from $\exp(e_i)$.

2.2.2. The Unbiased Load Estimate by the RCM Using IS

Now, we statistically explain that the load estimate by the RCM using IS is an unbiased estimator. IS is an efficient Monte Carlo integration method (e.g., Hammersley & Handscomb, 1964) that uses the probability distribution function (pdf) $g(t)$ to integrate $l(t)$ from t_1 to t_2:

$$L_{\mathrm{TRUE}} = \int_{t_1}^{t_2} l(t)dt = \int_{t_1}^{t_2}\frac{l(t)}{g(t)}g(t)dt \tag{7}$$

Equation 7 itself is identical regardless of the choice of $g(t)$. As $g(t)$ is a pdf, the integration of $g(t)$ over $[t_1, t_2]$ equals unity. Usually, $g(t)$ is defined as proportional to $l(t)$ in IS to achieve an efficient calculation. Notably, the support of $g(t)$ must include the support of $l(t)$ in IS (Robert & Casella, 2010): otherwise, Equation 7 gives a biased estimate due to the short support of l by g. Because we cannot know all the loading rates during $[t_1, t_2]$ in general, we should use the estimated loading rate (ELR) $\hat{l}(t)$, an approximate function of $l(t)$, instead of $l(t)$ to define $g(t)$:

$$g(t) = \frac{\hat{l}(t)}{\int_{t_1}^{t_2}\hat{l}(t)dt} \tag{8}$$

Sampling controlled by $g(t)$ is identical to sampling with a probability that is proportional to the size of an approximated or expected value (Valentine et al., 1984). Using $d\eta = g(t)dt$, we can rewrite Equation 7 as follows:

$$L_{\mathrm{TRUE}} = \left(\int_{t_1}^{t_2}\hat{l}(t)dt\right) \times \left[\int_0^1\frac{l(\eta)}{\hat{l}(\eta)}d\eta\right] \tag{9}$$

Using PPS samples based on a sampling pdf g, Equation 9 is discretized as the load estimate $\hat{L}_{IS}$:

$$\hat{L}_{IS} = \left(\sum_{i=1}^{N} \hat{l}_i\right) \times \left(\frac{1}{n_{\text{pps}}} \sum_{j=1}^{n_{\text{pps}}} \frac{l_{\eta j}}{\hat{l}_{\eta j}}\right) \tag{10}$$

where N is the population size and n_{pps} is the number of PPS samples. When $l_\eta/\hat{l}_\eta$ are nearly constant, that is, $\hat{l}(t)$ can approximate $l(t)$ well, Equation 10 can give a good approximation of L_{TRUE} or L_{ref} from small samples. Even when $\hat{l}(t)$ approximates $l(t)$ poorly, Equation 10 can still provide a good approximation of L_{TRUE} by increasing n_{pps}; therefore, $\hat{L}_{IS}$ is a consistent estimate.

When we approximate $l(t)$ using an RC as $\hat{l}(t) = \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}})$, the expectation of $\hat{L}_{IS}$ is written as follows:

$$E\left[\hat{L}_{IS}\right] = \left[\sum_{i=1}^{N} \exp\left(\mathbf{x}_i\hat{\boldsymbol{\beta}}\right)\right] \times E_g\left[\frac{l_\eta}{\hat{l}_\eta}\right] \tag{11}$$

where $E_g[\cdot]$ denotes the expectation of $\cdot$ under sampling pdf g. Since PPS samples are collected under g defined by Equation 8, the second factor on the right-hand side of Equation 11 becomes the following equation:

$$E_g\left[\frac{l_\eta}{\hat{l}_\eta}\right] = \int_0^1 \frac{l_\eta}{\hat{l}_\eta} d\eta = \int_{t_1}^{t_2} \frac{l(t)}{\hat{l}(t)} \times \frac{\hat{l}(t)}{\int_{t_1}^{t_2} \hat{l}(t)dt} dt = \frac{\int_{t_1}^{t_2} l(t)dt}{\int_{t_1}^{t_2} \hat{l}(t)dt} \cong \frac{\sum_{i=1}^{N} l_i}{\sum_{i=1}^{N} \hat{l}_i} \tag{12}$$

When n_{pps} is small and $\hat{\boldsymbol{\beta}}$ has a large deviation from $\boldsymbol{\beta}_m$, that is, $\hat{l}(t)$ approximates $l(t)$ poorly, Equation 12 does not hold true. That is, the quality of $\hat{l}(t)$ does affect the unbiasedness of load estimates when the sample size n_{pps} is small. On the other hand, Equation 12 would hold true when $\hat{\boldsymbol{\beta}}$ is determined by the OLS method. From Equations 11 and 12,

$$E\left[\hat{L}_{IS}\right] \cong \sum_{j=1}^{N} l_j = L_{\text{TRUE}} \tag{13}$$

From Equations 10 to 13, $\hat{L}_{IS}$ by Equation 10 is an unbiased estimator of L_{TRUE}. It must be noted that we made no assumption about the distributions of $\mathbf{X}_N$ or $\mathbf{e}_N$.

To investigate the relationship between $\hat{L}_{IS}$ and the load estimate by the ordinary RCM, we transform Equation 10 as follows:

$$\hat{L}_{IS} = \left[\sum_{i=1}^{N} \exp\left(\mathbf{x}_i\hat{\boldsymbol{\beta}}\right)\right] \times \left[\frac{1}{n_{\text{pps}}} \sum_{j=1}^{n_{\text{pps}}} \frac{\exp\left(\mathbf{x}_{\eta j}\hat{\boldsymbol{\beta}} + \delta_{\eta j}\right)}{\exp\left(\mathbf{x}_{\eta j}\hat{\boldsymbol{\beta}}\right)}\right] = \left[\sum_{i=1}^{N} \exp\left(\mathbf{x}_i\hat{\boldsymbol{\beta}}\right)\right] \times \frac{1}{n_{pps}} \sum_{j=1}^{n_{\text{pps}}} \exp \delta_{\eta j} \tag{14}$$

Equation 14 indicates that the smearing estimator (Equation 5, Duan, 1983; Koch & Smillie, 1986) based on PPS samples is unbiased; it is PPS sampling that provides unbiased load estimates. This equation is also identical to the load estimates by the SALT method using an RC as an auxiliary function of $l(t)$. From the above explanation, PPS or SALT sampling must comply with g defined by Equation 8 for unbiased estimation; hence, we should not interpret PPS or SALT sampling as an efficient sampling method weighted toward high flows. For example, the procedure employed in piecewise SALT (Thomas, 1988a, 1989) or later SALT application (Thomas, 1988b), in which the auxiliary variables were not based on $\hat{l}(t)$, causes bias in load estimates. To avoid such misunderstandings related to g, we refer to our proposed LEM as the RCM using IS instead of the SALT method.

2.2.3. An Implementation Method of PPS Sampling

Next, we consider an implementation method of PPS sampling required for the RCM using IS. Originally, Thomas (1985) introduced the SALT method as a realization method of PPS sampling at WQ monitoring

sites and employed the inverse function method based on uniform random numbers. That is, a sample was chosen at the location of a uniform random number on the cumulative ELR axis during $[t_1, t_2]$. This method requires knowing the total sum of the ELRs in advance. This sampling method uses uniform random numbers, the total sum, and a sophisticated automatic sampler that can manage SALT sampling. Such requirements have prevented PPS sampling from being adopted at monitoring sites. In this paper, we also employed the inverse function method for PPS sampling to evaluate the effectiveness of the RCM using IS despite its difficulty in applying to WQ monitoring sites. Another implementation method of PPS sampling at monitoring sites without using uniform random numbers will be discussed later in Section 4.2. The inverse function method also requires RC parameter values in advance of PPS sampling to calculate the cumulative ELRs at a given time during the target period. Namely, PPS sampling in the RCM using IS requires two-phase sampling, preliminary sampling or first-phase sampling to estimate RC parameters and inverse sampling itself as second-phase sampling.

2.3. The Confidence Interval of the Load Estimate by the RCM Using IS

2.3.1. Nonparametric Interval Estimation of the Load Estimates by the RCM Using IS

As mentioned earlier, although there have been many studies that discussed the uncertainty of load estimates using parametric CIs, few studies (e.g., Vigiak & Bende-Michl, 2013) have tested the performance of the prediction intervals or credible intervals of load estimates by comparing the probability or empirical coverage with which the constructed intervals can capture the reference or true load value with a prespecified significance level. Furthermore, a distribution of load estimates (not loading rates) has not been well understood, and the approximation of that distribution by a normal distribution based on the central limit theory requires large samples. Although other indices of the uncertainty of load estimates have been used, such as bias and variance (e.g., Walling & Webb, 1981), standard deviation (e.g., Richards & Holloway; 1987), or mean square error (e.g., Preston et al., 1989), these indices would not be useful without knowing the distribution of the estimates. We can conclude that we should use a nonparametric method for the uncertainty evaluation of load estimates, and the performance of the CIs must be tested using a reliable reference value. Therefore, we employed the bootstrap CI in this paper as the nonparametric CI of a load estimate $\hat{L}_{IS}$.

Here, we explain the method for constructing the CI of the load estimate based on the RCM using IS. Configuring the CI of the second factor on the rightmost side of Equation 14 leads directly to the interval estimation of $\hat{L}_{IS}$. Because $\delta_i = \mathbf{x}_i \times (\boldsymbol{\beta}_m - \hat{\boldsymbol{\beta}}) + e_i$ from Section 2.2.1, the distributions of both $\hat{\boldsymbol{\beta}}$ and e_i determine the distribution of δ_i. If we can assume normal distributions for both δ_i and e_i, we should use a parametric CI of the mean of n lognormally distributed random numbers, such as Land's H-statistics (Land, 1975; Singh et al., 1997; U.S. EPA, 2002) or generalized CIs (Krishnamoorthy & Mathew, 2003; Weerahandi, 1995). However, it is difficult to practically judge from small samples whether residuals δ_i follow a normal distribution or not because the goodness-of-fit statistics provide only the probability for incompliance with the distribution assumed. As a result, we adopted the bootstrap-t CI (Efron & Tibshirani, 1993) to construct the CI of a load estimate because of its higher coverage with small sample sizes. Note that bootstrap-t CIs will provide a lower coverage relative to the prespecified significance level for a mean of a finite number of lognormal variables when the sample size is small and the variance in residuals is large (Owen, 1988). We constructed and evaluated the central 95% CIs of load estimates in this paper.

The procedure for constructing the central $(1 - p) \times 100\%$ bootstrap-t CI is outlined as follows. For convenience of explanation, x_i ($i = 1$ to n) denotes exp (δ_i), and $\overline{x}$ denotes the sample mean of exp (δ_i). In the first step, the sample mean $\overline{x}$ and the jackknife estimate of the variance in the mean s^2 are calculated as $\Sigma x_i/n$ and $\Sigma(\overline{x}_{Ji} - \overline{x})^2/(n - 1)$, respectively, where $\overline{x}_{Ji}$ is the jackknife estimate of $\overline{x}$ calculated from the $(n - 1)$ samples other than the ith sample. In the next step, B sets of n bootstrap-replicated samples are extracted with replacement from the sample population of size n (B is set at 2,000 in this paper). Then, the bootstrap sample mean x^* and the jackknife estimate of the variance in the bootstrap sample mean s^{*2} are calculated for each bootstrap sample set. Using $\overline{x^*}$ and s^{*2}, the approximate pivot quantity t^* is defined

as the pivot statistic $\left(\overline{x^*} - \overline{x}\right)/s^*$ for each bootstrap sample set, and B sets of t^* are sorted in ascending order. Finally, the upper and lower confidence limits for the two-sided $(1-p)\times 100\%$ CI are calculated as $\overline{x} - s^2 \times t^*_{Bp/2}$ and $\overline{x} - s^2 \times t^*_{B(1-p/2)}$, respectively, where the ith order of t^* is denoted as t^*_i.

2.3.2. Limitation of the PPS Sample Size n_{pps}

In calculating $\hat{L}_{IS}$, the duplicated samples in the PPS samples cause bias. Interval estimation of $\hat{L}_{IS}$ comprises the interval estimation of the mean of a finite number of residuals of PPS samples extracted from the population. Hence, when duplicated samples are significant in the PPS samples extracted, the statistical properties of the distribution of the residuals of the PPS samples differ from those of the residuals of the population distribution. In this case, the resulting CIs constructed from PPS samples, including duplicated samples, cannot give proper coverage. In particular, we can hardly ignore this duplication problem in PPS sampling because PPS sampling tends to extract the sample with a large ELR. The avoidance of duplicated samples can be achieved by limiting the PPS sample size n_{pps}. For this purpose, we introduced the sample size $n_{\max}$ as the maximum PPS sample size:

$$n_{\max} = \mathrm{int}\left(\frac{\sum_{i=1}^{N}\hat{l}_i}{\hat{l}_{\max}}\right) \tag{15}$$

where $\hat{l}_{\max}$ is the maximum ELR during the target period. From Equation 15, $n_{\max}$ depends on the RC parameter $\hat{\boldsymbol{\beta}}$ and explanatory variables. In this paper, when the planned PPS sample size was greater than $n_{\max}$, the PPS sample size was again set at $n_{\max,}$ and only independent PPS samples were used in the load estimation.

Making PPS samples independent and limiting the sample size to $n_{\max}$ often force the resulting PPS sample size to be much smaller than the planned sample size; hence, the coverage of CIs may also be smaller than expected because of the coverage property of bootstrap-t CIs. When the PPS sample size is smaller than 10, the lower limit of bootstrap CI sometimes becomes less than zero. In this case, we set the lower limit of the CI at zero. In addition, when the resulting sample size is less than four, the lower limit often becomes minus infinity (Owen, 1988). To avoid the lower limit of minus infinity, we set the minimum PPS sample size to four. Therefore, when $n_{\max}$ is less than four, load estimation was not carried out in this paper.

2.4. Evaluation of Point and Interval Estimations by the RCM Using IS

2.4.1. RC Models Used in Estimation

We investigated the performance of load estimation using RCs in the Loadest model (Runkel et al., 2004), namely, Model 1 (two-parameter power law RC, Equation 16), Model 6 and Model 7 (five-parameter RCs, Equations 17 and 18, respectively), Model 9 (seven-parameter RC, Equation 19), and the RC model with the minimum AIC (Akaike information criterion, Akaike, 1974) value selected from Model 1 to Model 9 (Auto):

$$\ln\hat{l}_i = \beta_0 + \beta_1 \ln\frac{q_i}{q^*} \tag{16}$$

$$\ln\hat{l}_i = \beta_0 + \beta_1 \ln\frac{q_i}{q^*} + \beta_2\left(\ln\frac{q_i}{q^*}\right)^2 + \beta_3 \sin 2\pi T_i + \beta_4 \cos 2\pi T_i \tag{17}$$

$$\ln\hat{l}_i = \beta_0 + \beta_1 \ln\frac{q_i}{q^*} + \beta_3 \sin 2\pi T_i + \beta_4 \cos 2\pi T_i + \beta_5\left(T_i - T^*\right) \tag{18}$$

$$\ln \hat{l}_i = \beta_0 + \beta_1 \ln \frac{q_i}{q^*} + \beta_2 \left(\ln \frac{q_i}{q^*} \right)^2 + \beta_3 \sin 2\pi T_i + \beta_4 \cos 2\pi T_i + \beta_5 \left(T_i - T^* \right) + \beta_6 \left(T_i - T^* \right)^2 \quad (19)$$

$$\ln q^* = \overline{\ln q} + \frac{\sum_{i=1}^{n} \left(\ln q_i - \overline{\ln q} \right)^3}{2 \sum_{i=1}^{n} \left(\ln q_i - \overline{\ln q} \right)^2}, \quad T^* = \overline{T} + \frac{\sum_{i=1}^{n} \left(T_i - \overline{T} \right)^3}{2 \sum_{i=1}^{n} \left(T_i - \overline{T} \right)^2} \quad (20)$$

where n is the sample size, $\beta_0 \sim \beta_6$ are RC parameters (partial regression coefficients), T_i is the decimal time of the sampling, $\overline{T}$ is the mean of T_i, $\ln q_i$ and $\overline{\ln q}$ are the natural logarithm of q_i and its mean, respectively, and both $\ln q^*$ and T^* are introduced to remove the collinearity between the first- and second-order terms of discharge and time, respectively. Both Model 6 and Model 9 include the quadratic terms of discharge or decimal time as explanatory variables to express the nonlinearity of the RCs in log space. All the RC parameter values of these models, summary statistics of the residuals, and the results of Breusch-Pagan tests (Breusch & Pagan, 1979) of heteroscedasticity in residuals for the population of each data set are shown in Tables S1a–S1j.

2.4.2. Evaluation of Point and Interval Estimates by the RCM Using IS

Using the above RCs, we evaluate the bias of point estimates and the performance of interval estimation by the RCM using IS. As described in Section 2.2.2, when the PPS sample size n_{pps} is small and the deviation between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_m$ is large, that is, when $\hat{l}(t)$ does not approximate $l(t)$ well, $\hat{L}_{IS}$ would have a bias. Such a condition occurs when the RC parameters in first-phase sampling have a bias relative to those in second-phase sampling or when the RC parameters for a single year were used for PPS sampling during the following, much longer period. In practice, the suppositional conditions of the RCM using IS where there is no bias between the RC parameters in first-phase sampling and those in second-phase sampling are hardly satisfied; that is, these parameters are more or less different (practical conditions). This consideration made us adopt two types of evaluations: one is under suppositional condition to evaluate the intrinsic performance of the proposed method, and the other is under practical condition to test the potential performance of the method at the monitoring sites. In addition, the five types of the Loadest model listed above have been used to evaluate the effect of RC model selection on load estimation. In each evaluation, we extracted 2,000 sets of PPS samples from the population of each data set using the inverse function method and evaluated bias and the coverage of the CIs.

To evaluate the bias of load estimates, we used the following index, pBIAS (%):

$$\mathrm{pBIAS} = \left(\frac{1}{M} \sum_{i=1}^{M} \hat{L}_i - L_{\mathrm{ref}} \right) / L_{\mathrm{ref}} \times 100 \quad (21)$$

where M is the number of load estimates by Monte Carlo simulations and $\hat{L}_i$ is the load estimate. The performance of interval estimation was evaluated by the proximity of the coverage of the CIs to 95%. In calculating load estimates and sampling simulation, we used FORTRAN codes. Pseudorandom numbers were generated using the Mersenne Twister algorithm whose period is long enough for our Monte Carlo evaluation (Matsumoto & Nishimura, 1998).

2.4.2.1. Evaluation Procedure for Ideal Conditions

For the evaluation under ideal but suppositional conditions in which $\hat{\boldsymbol{\beta}}$ is an unbiased estimate of $\boldsymbol{\beta}_m$, first-phase sampling was carried out during the same period as second-phase sampling. This is because samples for the determination of RC parameters extracted randomly from the target period of the load estimation can provide unbiased RC parameters. In this evaluation, 20 samples in first-stage sampling were extracted randomly from the period of 2009 to 2011 for the GZ-Cl, GZ-K, and GZ-Na data sets and from the period of 2012 to 2014 for GZ-SS. Using the RC parameters based on these samples, PPS samples were collected by second-phase sampling, whose three sizes were 71–83, 36–42, and 16–19 for the solute

data sets in the Gozyo catchment (GZ-Cl, GZ-K, and GZ-Na) and 83, 42, and 19 for GZ-SS. These three sizes of PPS samples were equal to those obtained by weekly, fortnightly, and monthly systematic sampling during the target periods.

In the evaluation based on the US data sets under ideal conditions, we evaluated both annual and long-term load estimates, such as 5-year or decadal loads (Table S2). We evaluated the performance of the proposed method for long-term load estimates because the deteriorated performance was expected under a nonstationary RC relationship due to anthropogenic activities or climate changes during the long-term period. In short, a longer target period can cause a worse approximation of $l(t)$ by $\hat{l}(t)$, larger estimation bias, and poorer coverage. In evaluating the performance of the method for long-term loads, we gathered 12 × (the number of years in the target period) samples randomly from the target periods in the first phase and collected PPS samples in the second phase, whose planned three sizes were 488–521, 113–121, and 37–40 for decadal loads in the MK-N, RC-TP1, RC-TP2, RC-TP3, and SK-SS data sets and 239, 55, and 18 for the 5-year load in VM-N. The sample sizes of the PPS samples were equivalent to those by weekly, monthly, and seasonally systematic sampling during the target periods. We also evaluated the performance of the method for annual loads of the second year in the data sets. In these evaluations of annual loads, 26 samples were also collected randomly from the second year in the data sets in the first phase, and the PPS samples were collected in the second phase, whose planned sizes were 52, 26, and 12, which were equal to weekly, fortnightly, and monthly systematic sample sizes, respectively.

2.4.2.2. Evaluation Procedure for Practical Conditions

For the evaluation under practical conditions in which $\hat{\boldsymbol{\beta}}$ could be a biased estimate of $\boldsymbol{\beta}_m$, we also evaluated both annual and long-term load estimates using the US data sets. Under practical conditions, that is, in load estimation by the RCM using IS at monitoring sites, we must carry out PPS sampling in the target period using RC parameters based on the first phase samples collected in the period prior to the target period. To replicate such conditions, we randomly extracted 26 samples in the first phase from the water year (WY) 2008, WY1988, WY1998, WY2008, WY2008, and WY2002 for the MK-N, RC-TP1, RC-TP2, RC-TP3, SK-SS, and VM-N data sets, respectively, and carried out PPS sampling from the target period following the first phase WY. The planned PPS sample sizes are equal to the sizes tested in the evaluations under ideal conditions.

3. Results

3.1. Performance of Load Estimation by the RCM Using IS for Ideal Conditions

3.1.1. Bias of the Estimates and Coverage of the CIs

Figure 1a shows pBIAS values, and Figure 1b shows the coverage of the CIs of load estimates by the RCM using IS for all the data sets. The medium gray represents the zero pBIAS value (unbiased results) in Figure 1a and the light gray represents the 95% coverage value in Figure 1b. The negatively larger pBIAS values or lower coverage shift to the darker side, and positively larger pBIAS values shift to the brighter side. An open box in Figure 1a indicates an absolute pBIAS value larger than or equal to 10%, and an open box in Figure 1b indicates coverage of less than or equal to 90%. The results in Figures 1a and 1b without an open box indicate almost unbiased load estimates and the appropriate coverage of CIs, respectively. The upper half of Figure 1 shows the results for ideal conditions where the periods of the first- and second-phase sampling are the same, and the lower half gives the result under practical conditions, where the period of first-phase sampling is prior to second-phase sampling. The detailed values are listed in Tables S3–S8.

Before we evaluate the estimation results, we summarize the statistical properties of the data sets. Tables S1a–S1j show the skewness and kurtosis of the residuals whose values for a normal distribution are zero and 3.0, respectively. For the heteroscedasticity of the residuals, the p-values for the null hypothesis (i.e., homoscedasticity) by the Breusch-Pagan test are also given in these tables. The p-values less than 0.05 indicate heteroscedasticity in the residuals at a 0.05 level of significance. From these tables, hetero-

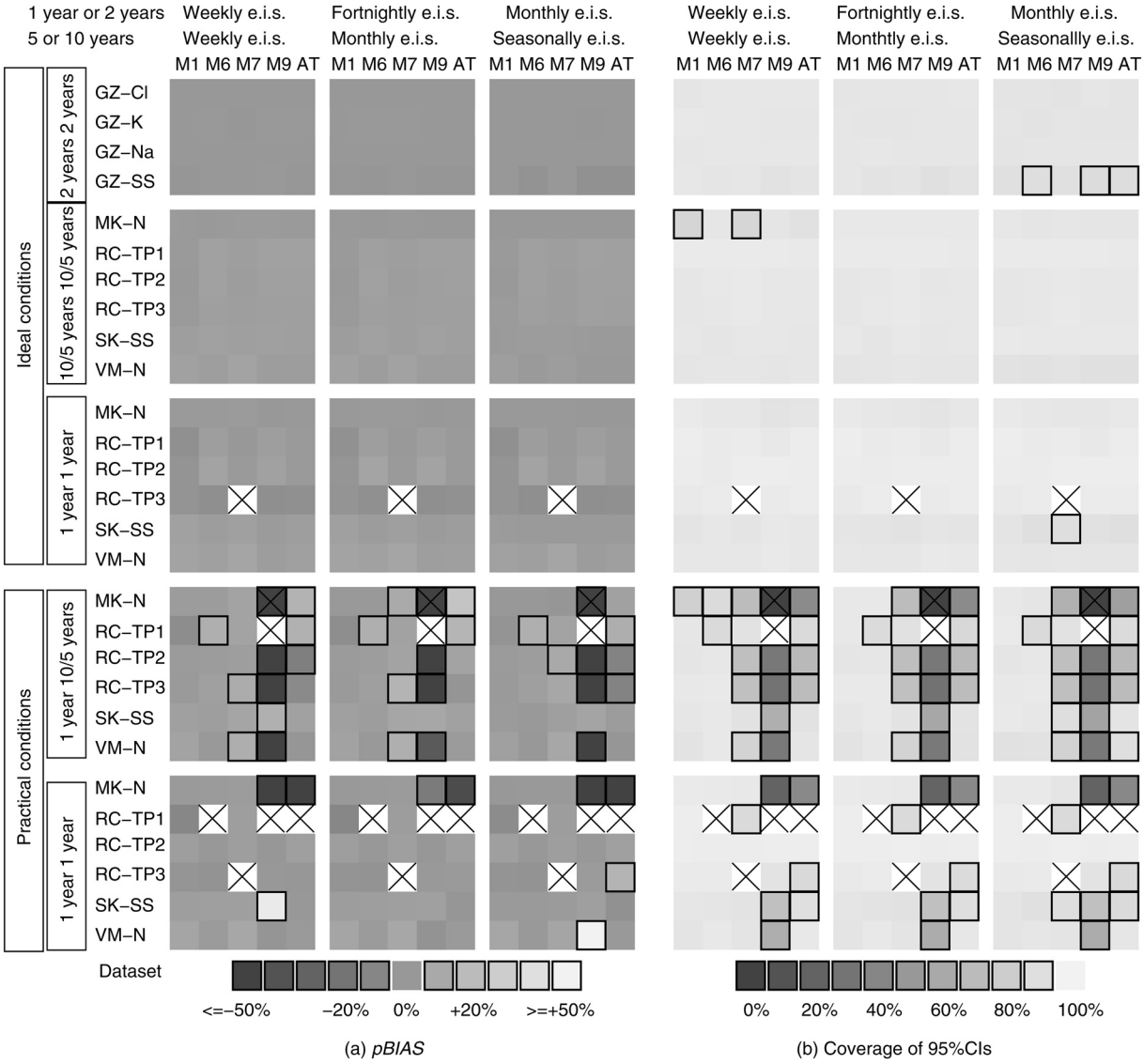

Figure 1. pBIAS and coverage of 95% confidence intervals (CIs) of the load estimates by the rating curve method (RCM) using IS. M1, M6, M7, M9, and AT represent the load estimates using Model 1, Model 6, Model 7, Model 9, and the autoselected model based on the minimum AIC value, respectively. An open frame indicates values ≤−10% or ≥10% in pBIAS and values less than or equal to 90% in coverage. The abbreviation "e.i.s." means "equivalent in size." The cross symbol denotes "no evaluation" due to a small n_{max}.

scedasticity in the residuals was found in all Gozyo data sets and in the MK-N, RC-TP2, SK-SS, and VM-N data sets, and higher kurtosis was found in solute data sets in the Gozyo catchment and the MK-N data set. The tested data sets do not satisfy the traditional assumption of linear regression.

In the ideal conditions tested, most of the pBIAS values lie within ±5% and do not exceed ±10% for any sample size, any RC model, any watershed, or any WQ parameter. The pBIAS values are almost zero in GZ-Cl, GZ-K, and GZ-Na, where slope parameter β_1 in Model 1 is nearly equal to unity and where regression residuals have a small variance (Tables S1a–S1c). In the Gozyo data sets, although the coverage of CIs for GZ-SS ranges from 88% to 90%, the coverages for other parameters have values comparable to 95% (90%–98%). The lower coverage in GZ-SS is attributed to the ability of bootstrap-t CIs to construct CIs for the mean of finite samples from a lognormal distribution with a large variance (Owen, 1988). In practice, the distribution of the regression residuals of the GZ-SS population by Model 1 has a skewness of −0.59, a kurtosis of 3.04, and a variance of 1.33 (Table S1d), and these values of skewness and kur-

tosis are comparable with those of a normal distribution. On the other hand, although the regression residuals of the population of RC-TP1, RC-TP2, RC-TP3, SK-SS, and VM-N by Model 1 have values of skewness and kurtosis that are more comparable with those of a normal distribution (Tables S1f–S1j), their variances that are smaller than those of GZ-SS place the coverage closer to 95%. For the coverage of the long-term load estimates in the upper half of Figure 1b, which are comparable to the results by the ordinary RCM in Figure S1b, most of the results show coverage comparable to 95%, except for the low coverage (84%–85%) shown in the results for Model 1 and Model 7 at weekly sample sizes ($n = 505$) in MK-N. A large skewness of 2.28 and large kurtosis of 13.3 of the distribution of the regression residuals of MK-N by Model 1 (Table S1e) would be a major reason for these low coverage. In summary, the RCM using IS can provide unbiased load estimates and proper coverage of CIs as long as $\hat{l}(t)$ is the unbiased estimate of $l(t)$.

3.1.2. Effect of RC Model Selection

As shown in Figure 1, selection of the RC model affects neither the unbiasedness of the estimates nor the appropriateness of the CIs coverage under ideal conditions. The cause of bias in load estimation by the ordinary RCM has been attributed to the violation of the traditional assumptions of the linear regression model: normality of errors with zero mean, no systematic errors by the model, homoscedasticity of errors, and the absence of autocorrelation in errors. Koch and Smillie (1986) attributed the bias to nonnormality of residuals and both Stenback et al. (2011) and Hirsch (2014) focused on heteroscedasticity of residuals as the cause of bias. Nevertheless, the RC using IS can provide unbiased load estimates even for the observation data sets with heteroscedastic RC residuals shown in Tables S1a–S1j. These unbiased results are contrasted with the biased load estimates by the ordinary RCM based on random sampling shown in Text S1, Figure S1, and Tables S9–S12. It should be noted that the load estimates by the smearing estimator L_{SMR} (defined in Text S1) use the same calculation as those by the RCM using IS; their difference lies only in the difference between random and PPS sampling. The bias of L_{SMR} for US data sets is substantial in Figure S1, whereas the results by the RCM using IS are almost unbiased.

Furthermore, in these results, poorly determined RC parameters compared to the population values are especially expected in Model 6 and Model 9 based on small-size random samples. In practice, the load estimates for US data sets by the ordinary RCM (Table S11) have moderate (approximately 50%) to large (over 100%) pBIAS values as reported by Lee et al. (2016). In contrast to these biased results by the ordinary RCM, similarly poorly determined RC parameters were used in PPS sampling, and the RCM using IS could calculate unbiased load estimates for all the models evaluated. Moreover, the RCM using IS can also provide unbiased estimates using both the simple RC model and more sophisticated model that can reduce systematic errors in RC residuals. Namely, this unbiasedness by the RCM using IS arises from PPS sampling and does not much deteriorate from nonnormality or heteroscedasticity of RC residuals. The dominant cause of bias lies in the ignorance of the logical connection between the sampling strategy and LCM (Thomas & Lewis, 1993). In regard to the load estimation for long-term records, pBIAS values and the coverage of load estimates for RC-TP1, RC-TP2, and RC-TP3 have similar values, but the resulting PPS sample sizes are different between them.

3.1.3. Effect of Resulting PPS Sample Size

We also demonstrated the distributions of point estimates, lower limits of CIs (*LL*s), and upper limits of CIs (*UL*s) of load estimates for VM-N and SK-SS in ideal but suppositional conditions in Figure 2. In Figure 2, the load estimates by the RCM using IS are almost unbiased regardless of sample size or the RC model employed and the coverages range from 90% to 95%. In regard to the effect of sample size on the uncertainty of load estimates, the widths of CIs are narrower since n_{pps} is larger, as shown in Figure 2. The sample size and an approximation quality of $\hat{l}$ do not affect the unbiasedness or coverage of CIs but affect the variance or precision of load estimates such as the width of CIs. To acquire more precise or less uncertain load estimates, more PPS samples are required, as shown in Figure 2.

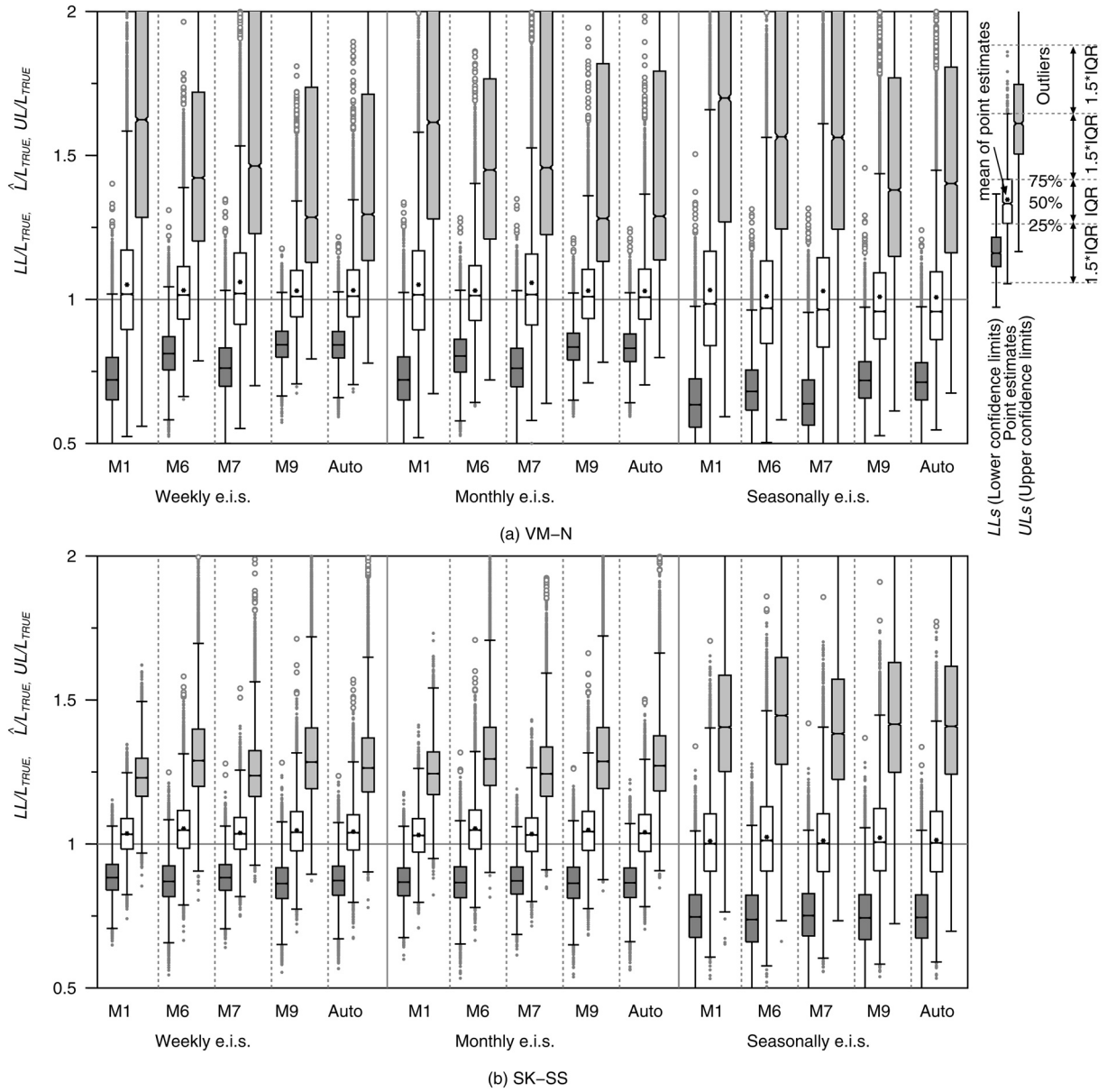

Figure 2. Box and whisker plots of the upper limits and lower limits of the two-sided 95% CIs and point estimates for decadal/5-year load estimates by the RCM using IS under ideal conditions. M1, M6, M7, M9, and Auto represent the load estimates using Model 1, Model 6, Model 7, Model 9, and the autoselected model based on minimum AIC, respectively. The abbreviation "e.i.s." means "equivalent in size.".

3.2. Performance of Load Estimation by the RCM Using IS for Practical Cases

3.2.1. Bias of the Estimates, Coverage of the CIs, and Effect of PPS Sample Size on the Performance

For practical conditions where the period of first-phase sampling is prior to the period of second-phase sampling, the detailed values of the results are also listed in Tables S5–S8. Under practical conditions, the RC parameters used in PPS sampling were not necessarily unbiased estimates of the population parameters for the target period. For evaluation under practical conditions, we used six US data sets because the observation length of the Gozyo data sets was not long enough. In the estimation using the RC models, except Model 1, we could not always perform load evaluations because of the limitation of the number of PPS samples by n_{max}, whose value is less than four. This small value of n_{max} in Equation 15 comes from an extremely large maximum value of ELR. This large maximum ELR was caused by the RC parameter values

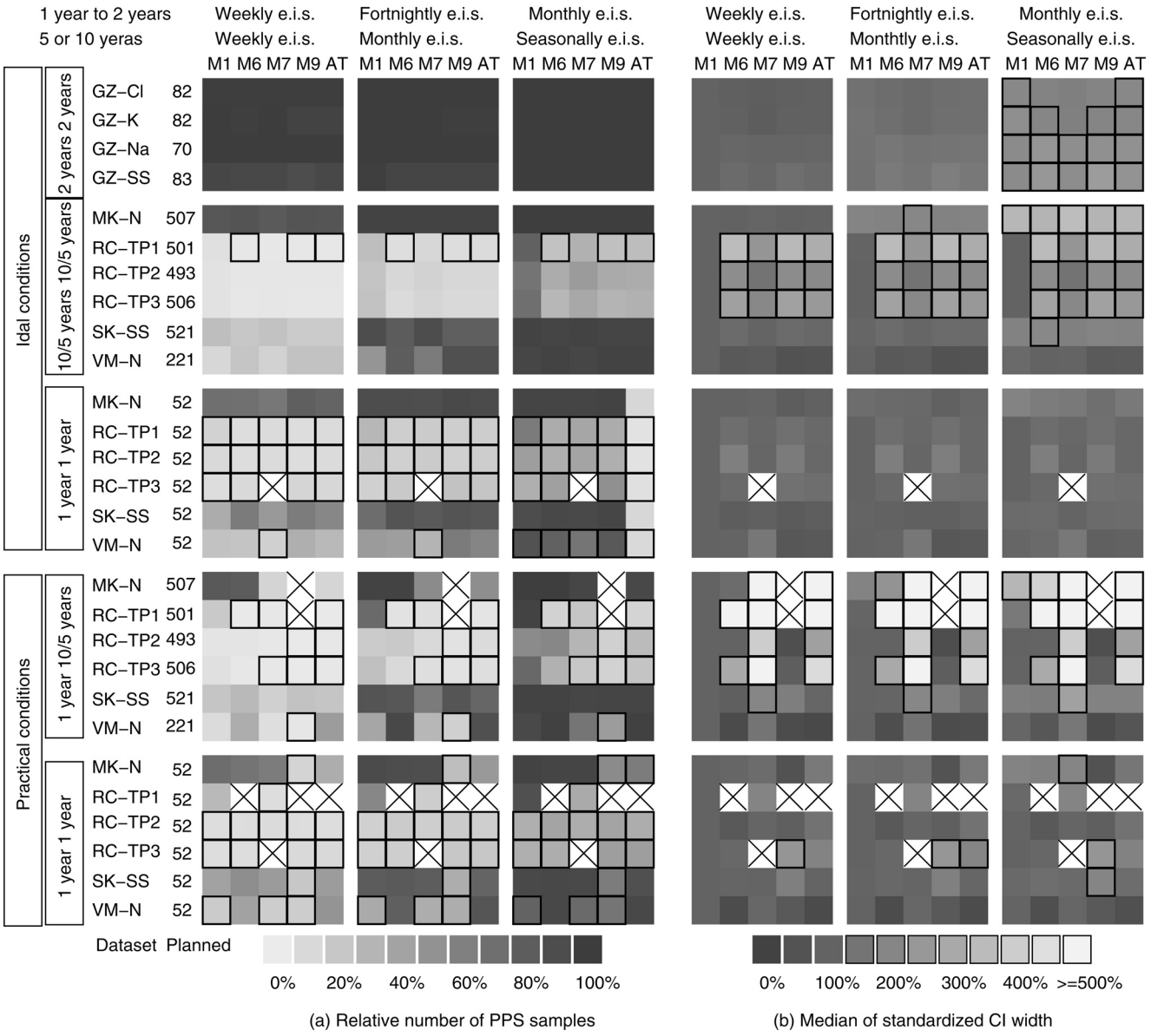

Figure 3. The relative numbers of proportional to size (PPS) samples to the planned sample numbers and the medians of relative CI width standardized by the median width of CIs using Model 1. M1, M6, M7, M9, and AT represent the load estimates using Model 1, Model 6, Model 7, Model 9, and the autoselected model based on the minimum AIC value, respectively. Open frames in (a) indicate PPS sample numbers ≤10. Open frames in (b) indicate medians of standardized CI widths greater than or equal to 200%. The abbreviation "e.i.s." means "equivalent in size." The cross symbol denotes "no evaluation" due to small n_{max}.

that deviated from those of the population parameter values. To illustrate the effect of n_{max} on the resulting PPS sample size, we summarize the ratio of the resulting PPS sample size to the planned size equal to the comparable systematic sampling, as shown in Figure 3a. In this figure, the open frames indicate the cases where the resulting PPS sample size is less than 10. We also listed the individual values of the resulting PPS sample sizes in Tables S13–S15.

The lower half of Figure 1a shows that the biases in load estimates by Model 1 tend to be smaller than those by the other RC models, and this tendency is more significant in long-term load estimations. The comparison of the results between the ideal and practical conditions indicates that the unbiasedness by the RCM using IS becomes impaired when the RC parameters in the first phase samples cannot approximate the RC parameters in the target period well. Furthermore, this figure also suggests that the load estimates based

on Model 1 are more robust than those based on the other RC models because the RC models with more explanatory variables can inflate the uncertainty in the ELR compared with the model with minimum explanatory variables. The lower half of Figure 1b shows that long-term load estimation by the RCM using IS provides lower coverage than single-year estimation because of the poorer approximation of $l(t)$ by $\hat{l}(t)$ during the target period in long-term evaluations. From the results of the single-year estimation, the addition of quadratic terms in the explanatory variables of the RC models leads to lowered coverage. The comparison of the resulting PPS sample size under ideal conditions to those under practical conditions for six US data sets shows a reduction in sample size under practical conditions and this reduction is significant in Model 7, Model 9, and Auto, which have more explanatory variables. We also recognize that the reduction in sample size in the long-term load estimation is larger than that in the single-year estimation. This fact indicates that the uncertainty in RC parameters used in PPS sampling more severely affects the uncertainty in the load estimates in long-term estimations. The reduced number of resulting PPS samples by the limitation of n_{max} was significant in the load estimations by Model 7, Model 9, and Auto for RC-TP1, RC-TP2, RC-TP3, where the variance in the regression residuals is large and the slope parameter β_1 of Model 1 is much larger than 1.0. In these data sets, n_{max} became too small because the logarithms of their discharge data with large variance and skewness (see Table S16) resulted in a small number of PPS samples extracted from high flows contributing significantly to the total annual discharge. The pBIAS values and coverage of long-term load estimates are different between RC-TP1, RC-TP2, and RC-TP3 because of the biased approximation of $l(t)$ by $\hat{l}(t)$.

As shown in the difference between the results of ideal and those of practical conditions, unbiased load estimation by the RCM using IS requires the unbiased estimation of $l(t)$. Originally, the SALT method was developed as an alternative to PPS sampling from the population that could not be revisited (Thomas & Lewis, 1995), resulting in the use of an auxiliary variable $\hat{l}(t)$, that is, a proxy function of $l(t)$, and two-phase sampling. Practically, we can only retrospectively calculate the unbiased proxy function of $l(t)$ for the target period, and we cannot avoid using a more or less biased proxy function of $l(t)$ in load estimations. Accordingly, we must find a solution to avoid increased bias and lowered coverage caused by the biased $\hat{l}(t)$ in future applications of the RCM using IS at WQ monitoring sites.

3.2.2. Desirable Intervals of Discharge Data for Load Estimation by the RC Using IS

A daily interval of discharge in the tested US data sets led to small sample sizes of the resulting PPS samples, as shown in Tables S13–S15, despite their long target periods. The value of n_{max} defined by Equation 15 tends to be much smaller based on the daily data when the slope coefficient between the logarithms of loading rates and discharge is larger than unity. Since n_{max} is defined as the division of the total sum of ELRs for the target period by the maximum ELR, a shorter interval of discharge, such as an hourly or subhourly interval, would provide a smaller maximum ELR value and a relatively robust total sum of ELRs; therefore, a shorter interval would lead to a larger n_{max}. In turn, a larger n_{max} would result in unbiased load estimation. A daily interval of discharge data is also inadequately long for load estimation because neither discharge nor concentration is constant for 24 h in high flows, even for large watersheds as described in Section 2.1. For less uncertain load estimation, we should use a data interval that is short enough to express the changes in concentration or discharge.

4. Discussion

4.1. Effects of Model Selection on the Unbiasedness of the Load Estimates by the RCM Using IS

The results in Sections 3.1.2 and 3.2.1 indicate that the addition of explanatory variables to Model 1 in the RCM using IS tends to lower the coverage of the CIs, whereas it is considered that the complicated RC with rich explanatory variables can express seasonal variation, trends, or nonlinearity between concentration and discharge in log space can reduce bias and the uncertainty in load estimates. We now investigate the effect of the addition of explanatory variables on the uncertainty in load estimates. Figure 3b shows the median of the CI widths from 2,000 evaluations for 95% CIs, and its value is standardized by the median width of the CIs using Model 1 and expressed in percent. The individual median widths of the CIs standardized by L_{ref} are listed in Tables S17–S19. In Figure 3b and these tables, the median widths of CIs by Model 1 are

the narrowest or comparable to the narrowest among those of the tested models except VM-N and especially under ideal conditions; hence, the addition of explanatory variables does not necessarily reduce the uncertainty in load estimates.

The first reason for the increase in the uncertainty in estimates by adding explanatory variables to Model 1 comes from RC overfitting to the samples in first-phase sampling, or in other words, RC extrapolation in calculating load estimates. Since the RC parameters are determined to fit the samples well in the first phase, these parameters can inflate the uncertainty in ELR near or beyond the limit of the values of explanatory variables in these samples, resulting in an increase in the uncertainty in load estimates. This inflation is more significant when a complicated RC with rich explanatory variables is used (Hirsch, 2014; Slates et al., 2017). It is important to note that we cannot avoid the extrapolation of an RC when we use the RCM for load estimation. Although Gilroy et al. (1990) and Cohn (1995) recommended avoiding extrapolation by an RC for accurate load estimation, we can avoid extrapolation when the sample size is large enough or when we employ sampling strategies that can cover the possible range of explanatory variables for the target period. These solutions for avoiding extrapolation are difficult to achieve, particularly when a small sample size is desired. In load estimation, we should recognize the RCM as an extrapolation method (Littlewood, 1992; Walling & Webb, 1981).

The second reason for the increase in the uncertainty of estimates by adding explanatory variables is attributed to the small sample size of the resulting PPS samples under the limitation of small n_{max} caused by the extremely large ELR. From Tables S14 and S15 and from Figure 3a, we find the largest sample size mostly from Model 1 except VM-N. From Tables S18 and S19 and Figure 3b, we also find the narrowest median widths of the CIs mostly from Model 1 except VM-N. The addition of explanatory variables to Model 1 leads to a reduction in the resulting PPS sample size. Lee et al. (2019) reported that less accurate load estimates by the RCM were found in more variable loading conditions where the difference between the minimum and maximum loading rates is large, and this loading condition also increases the uncertainty in load estimates by the RCM using IS through reduction in the PPS sample size due to the limitation from the small n_{max} value. From the above considerations, we recommend using the traditional two-parameter power law RC, Model 1, in the application of the RCM using IS at monitoring sites unless more information on the appropriate RC model is acquired.

4.2. Implementation of the RCM Using IS for WQ Monitoring Sites

4.2.1. Practical PPS Sampling Method Without Using Uniform Random Numbers

4.2.1.1. PPS Sampling at Every Aliquot of Cumulative ELRs

The SALT method (Thomas, 1985) required using uniform random numbers in second-phase (PPS) sampling based on intelligent sampling equipment (Cohn et al., 1992; Thomas, 1989). This requirement hindered the application of the SALT method to WQ monitoring sites. In practice, based on Model 1, we can also conduct PPS sampling at monitoring sites using equally spaced numbers, which are restricted uniform random numbers, through the cumulative ELRs instead of using uniform random numbers. Using this method, we can collect PPS samples using automatic samplers available on the market (e.g., Teledyne Isco Inc., 6,712 portable sampler with its optional 720 submerged probe flow module) at the site where discharge correlates with the river water stage and where the RC parameters of Model 1 have been determined in advance. In this case, the real-time discharge value is calculated using the water stage measured by the probe, and the ELR is calculated using the RC. The ELR value is then added to the cumulative ELRs, and the PPS sample is collected at every predefined aliquot of cumulative ELRs. Henceforth, we call this PPS sampling method ΔL sampling. It is important to ensure the appropriateness of the discharge RC and the representativeness of the concentration at the sampling point in a river cross-section as the average concentration over the cross-section, for example, through calibration. From this point, ΔL sampling would be suitable for small streams rather than large rivers.

Next, we should investigate the difference between load estimates by ΔL sampling and those by inverse function sampling based on uniform random numbers, hereinafter called random PPS sampling. For this purpose, we calculate load estimates by the RCM using IS based on ΔL sampling for the GZ-Cl, GZ-K, GZ-Na, and GZ-SS data sets in the same procedures described in Section 2.4.2. We used Model 1 in this

Table 2
Relative Biases (pBIASs) and CI Coverage of 2-Year Load Estimates by the RCM Using IS With ΔL Sampling Based on the Observed Data

	pBIAS (%)				Coverage (%)			
Sample size	GZ-Cl	GZ-K	GZ-Na	GZ-SS	GZ-Cl	GZ-K	GZ-Na	GZ-SS
Weekly e.i.s.	0.0	0.0	0.0	0.2	100	100	99.9	98.8
Fortnightly e.i.s.	0.0	0.1	0.0	−0.2	100	99.9	98.1	95.9
Monthly e.i.s.	0.0	0.0	0.0	−0.2	99.7	99.7	98.4	95.6

Abbreviation: e.i.s., equivalent in size.

evaluation, and the results were compared with those of random PPS sampling shown in Table S3 for pBIAS values and in Table S4 for the coverage of CIs.

Table 2 shows the pBIAS values and CI coverage of load estimates based on ΔL sampling. Whereas the pBIAS values by ΔL sampling are as small as those by random PPS sampling, the coverages by ΔL sampling are higher than those by random PPS sampling and higher than the expected significance level of 95%. These higher coverages are less significant in GZ-SS, although these differences depend on the sample size and WQ parameters. In fact, these higher coverages are the results of autocorrelation between regression residuals. We will discuss the effects of the statistical properties of regression residuals on load estimates below.

4.2.1.2. Effect of Serially Correlated RC Residuals on Load Estimates by ΔL Sampling

The OLS method commonly used in RC regression poses four traditional assumptions on the regression residuals: normality with zero mean, absence of systematic errors, homoscedasticity, and independence. Among these four assumptions, incompliance with the first three assumptions would cause bias and lower the coverage of the load estimate CIs. Accordingly, incompliance with the last assumption, in other words, the presence of autocorrelation between residuals, would cause higher coverage for the CIs. Practically, we cannot expect independent regression residuals of the population in the RCM. We can usually recognize autocorrelation between regression residuals, as pointed out in many past studies: improvement in RC estimates accounting for serial correlation between residuals (Zhang & Hirsch, 2019), low uncertainty of load estimates due to autocorrelation (Zamyadi et al., 2007), more accurate interpolation methods based on serial correlation (Aulenbach, 2013), load estimation by a linear mixed model accounting for serial correlation (Lessels & Bishop, 2013; Slates et al., 2014, 2017), and load estimation using generalized RC considering serial correlation in residuals (Kuhnert et al., 2012; Wang et al., 2011). The application of simple models, such as Model 1, to high-frequency data inevitably causes autocorrelation between residuals because residuals cannot behave randomly in chronological sequences (Cochran, 1977). The presence of autocorrelation in residuals or loading rates would make the bias of load estimates based on interpolation methods (e.g., Walling & Webb, 1981), such as the period-weighted average method (e.g., Kronvang & Bruhn, 1996; Moatar & Meybeck, 2005) or composite method (Aulenbach, 2013; Aulenbach & Hooper, 2006; Aulenbach et al., 2016) small (but not controlled). Load estimates by these interpolation methods tend to have smaller biases with shorter observation intervals of WQ and discharge data, although the resulting bias depends on the coefficient of autocorrelation. Actually, the second or higher order autocorrelation between regression residuals of the population has been found based on the Yule-Walker method and the AIC in four Gozyo data sets.

To investigate the effect of autocorrelation between residuals on load estimates, we assume the first-order Markov process, that is, serial correlation, between the population residuals. The serially correlated residual e_i^* is expressed as follows:

$$e_i^* = \rho e^*_{i\text{-}1} + \sigma_c rnd_i \tag{22}$$

where ρ is an autocorrelation coefficient, σ_c is the standard deviation of residuals after removing serial correlation, and rnd_i is the ith normal random number from $N(0, 1)$. The synthetic loading rate l_{si} $(i = 1 \sim N)$

Table 3
Percent Biases (pBIASs) and CI Coverage of 2-Year Load Estimates by the RCM Using IS With ΔL Sampling Based on the Synthesized Data

	pBIAS (%)				Coverage (%)			
	NSC		SC		NSC		SC	
Sample size	Random	ΔL	Random	ΔL	Random	ΔL	Random	ΔL
Weekly e.i.s.	0.0	0.0	0.0	0.0	94.9	95.7	95.3	99.7
Fortnightly e.i.s.	0.0	0.0	0.0	0.0	94.8	95.1	95.1	99.8
Monthly e.i.s.	0.0	0.0	0.0	0.0	94.8	95.2	94.8	98.9

Note. NSC and SC correspond to the synthetic data without and with serially correlated residuals, respectively.
Abbreviation: e.i.s., equivalent in size.

with serially correlated residuals is expressed by substituting $e_i^* = \ln l_{si} - \beta_0 - \beta_1 \ln (q_i/q^*)$ and $e_{i-1}^* = \ln l_{si-1} - \beta_0 - \beta_1 \ln (q_{i-1}/q^*)$ into Equation 22 as follows:

$$ls_i = \exp\left[(1-\rho)\beta_0 + \beta_1\left(\ln\frac{q_i}{q^*} - \ln\frac{q_{i-1}}{q^*}\right) + \rho \ln ls_{i-1} + \sigma_c rnd_i\right] \tag{23}$$

Using Equation 23 and rnd_i, β_0, β_1, σ_c, ρ, and l_{s1}, l_{si} ($i = 2 \sim N$) is generated. Table S20 gives the values of β_{m0}, β_{m1}, σ_m^2, and ρ for the populations of GZ-Cl, GZ-K, GZ-Na, and GZ-SS based on the Cochrane-Orcutt method (Cochrane & Orcutt, 1949), together with the variance σ_c^2 of the residuals without serial correlation. We synthesized l_{si} for GZ-K using $\beta_0 = \beta_{m0} = -8.87$, $\beta_1 = \beta_{m1} = 0.883$, $\sigma_c^2 = 2.36 \times 10^{-2}$, $\rho = 0.998$, and rnd_1 and then calculated load estimates using (q_i, l_{si}). For comparison, the synthetic loading rate l_{di} without serial correlation was calculated using the following equation and was used for load estimation:

$$l_{di} = \exp\left(\beta_{m0} + \beta_{m1} \ln\frac{q_i}{q^*} + \sigma_m rnd_i\right) \tag{24}$$

Using these two synthetic data sets of (q_i, l_{si}) and (q_i, l_{di}), we calculated load estimates in the same procedures as described in Section 2.4.2, assuming ideal but suppositional conditions. The tested n_{pps} were 19, 41, and 82, equal to the sample size of monthly, fortnightly, and weekly systematic sampling, respectively. Table 3 gives the coverage of the CIs, and Figure 4 provides the distributions of *LL*s, point estimates, and *UL*s.

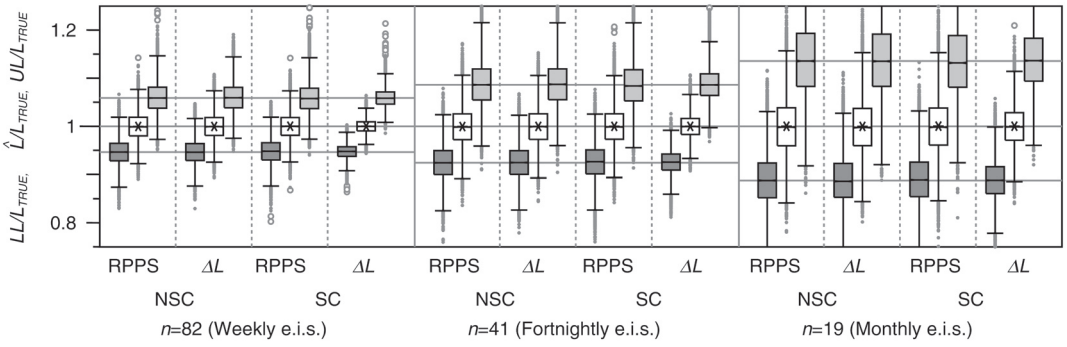

Figure 4. Box and whisker plots of the upper limits and lower limits of the two-sided 95% CIs and point estimates for load estimates by the RCM using IS based on the synthetic data. The abbreviation "e.i.s." means "equivalent in size." NSC and SC correspond to the synthetic data without and with serially correlated residuals, respectively. RPPS and ΔL indicate random PPS sampling and ΔL sampling, respectively. The style of the box and whisker plots is the same as that in Figure 2.

Table 3 indicates that (1) without serial correlation between residuals, the coverages are equal to the expected significance level regardless of the PPS sampling method used and (2) with serial correlation, PPS samples by random sampling provide the equivalent coverage to the expected significance level, whereas those by ΔL sampling give higher coverage. Figure 4 shows that the medians or averages of *LL*s, point estimates, and *UL*s are almost the same in all the cases evaluated, although the widths of the distributions (variance) of *LL*s, point estimates, and *UL*s of load estimates by ΔL sampling are narrower than those by random PPS sampling. Usually, random samples tend to provide larger variances than systematic or equally spaced samples (Corchan, 1977; Zamiyadi et al., 2007). These narrower CIs by ΔL sampling are caused by the similarity of the sample combination among the different PPS sample sets. This similarity arises from the property of PPS sampling that a larger ELR tends to have a higher probability of sampling. For example, PPS samples tend to be sampled from some large flood events, and the residuals in the same flood event tend to have similar values when strong serial correlation or autocorrelation exists in residuals. With autocorrelation between residuals, a similar PPS sample combination by ΔL sampling reduces the width of the distribution of estimates, and simultaneously, median values of *UL*s, point estimates, and *LL*s remain unbiased (stable). As a result, PPS samples by ΔL sampling with serially correlated residuals provide higher coverage relative to the expected significance level. Notably, we cannot estimate the quantitative properties of autocorrelation between residuals, such as σ_c^2 or ρ of the population, from infrequent samples. Hence, it is difficult to remove the effects of autocorrelation between residuals on load estimates. In summary, autocorrelation between residuals does not matter when we employ random PPS sampling and does not cause bias or lower coverage even when we use ΔL sampling.

4.2.2. Feasibility of the RCM Using IS in Field Monitoring Sites

4.2.2.1. Inefficiency in the RCM Using IS and Its Solution

As noted by Thomas (1985), the RCM using IS is an inefficient estimation method because one WQ parameter requires one sampling strategy corresponding to its RC parameters, that is, PPS sampling requires a different sampling schedule for a different RC parameter set. That is, one automatic sampler is required for one WQ parameter, not for one monitoring site. In practical situations, one automatic sampler should provide load estimates for multiple WQ parameters. The solution for this problem could be the application of the sampling/importance resampling (SIR) method (or importance resampling method, Rubin, 1987; Tanner, 2006) because this method enables the extraction of samples following some pdf from those following a different pdf. In addition, although we pointed out the bias caused by the biased RC parameters of the first phase samples compared with those of the population described in Section 3.2, the SIR method could also remove this bias. That is, by not using the RC parameters regressed from the first phase and instead using those from the second phase samples to calculate g used for resampling by the SIR method, we can remove this bias. Currently, the application of the RCM using IS to WQ monitoring sites is limited to small research catchments due to both inefficient sampling and poorly determined RC parameters, unless the solution by the SIR method is developed.

4.2.2.2. Effects of Missing Observations on Load Estimates

We should also discuss the effect of missing observations in the PPS samples on load estimates by the RCM using IS for application to the WQ monitoring sites. We suppose an interruption of PPS sampling for a certain period during the target period. Since missing observations of discharge make load estimation impossible for the target period, we do not need to discuss the effect of missing discharge observations. Here, we discuss the effect of missing WQ observations because of the malfunction of an automatic sampler or battery trouble on the premise of no missing discharge observations. In this case, load estimates by the RCM using IS would have bias because the support of g does not necessarily cover the support of $l(t)$. Even when the support of g covers the support of $l(t)$ despite the missing WQ observations, the resulting samples may not satisfy the PPS. For these reasons, we excluded missing observations in preparing the data sets in Section 2.1. In other words, to calculate accurate load estimates for the target period with missing WQ observations, new PPS samples must be extracted as a subset of the PPS samples collected so that the support

of g can cover the support of $l(t)$. Without a solution for missing observations, the RCM using IS cannot provide unbiased load estimates for the period with missing observations. The SIR method could also be a solution for this problem.

4.2.2.3. Effect of Observations Below the Detection Limit on Load Estimates

Type 1 censored data, which contain concentrations below the detection limit, are common in field observations for some WQ parameters. The application of the RCM using IS to censored data requires RC parameters regressed by the Tobit model (Tobin, 1958) in the first phase. In second-phase sampling, observations below the detection limit could be sampled from censored data. In this case, we can choose two options. First, the residuals of such undetected samples used in load estimation may be determined by resampling from residuals of samples over the detection limit or by estimation based on some parametric distribution applied to the detected-sample residuals. Second, the samples below the detection limit are treated as missing observations. This option would be preferable because it does not require any assumption of residuals for undetected samples, although the problem of applying the RCM using IS for the data with missing observations arises in turn. Without a solution for missing observations, the RCM using IS cannot provide unbiased load estimates even for censored data.

4.3. Implications for the Current WQ Monitoring Programs

As described in Section 3.1, the unbiased load estimation by the RCM using IS reveals that the cause of bias in load estimates by the ordinary RCM lies in improper sampling, except PPS sampling. For example, absolute values of the pBIASs of the estimates by the RCM using IS did not exceed 10%, except the results using Model 9 (Tables S3, S5 and S7). On the other hand, the absolute pBIASs of the smearing estimate by the ordinary RCM expressed as L_{SMR}, which has the same equation to calculate the load as the RCM using IS (except for the sampling method), often exceed 10% and sometimes 100% based on random sampling (Table S11). From these results, PPS sampling does lead to unbiased load estimation. Considering that random samples tend to collect low-flow samples in accordance with the flow distribution of the population, the comparison between L_{SMR} and $\hat{L}_{IS}$ demonstrates the importance of high-flow samples in unbiased load estimation. However, high-flow sampling should follow the PPS probability for this purpose.

Currently in the United States, some WQ monitoring programs employ a combination of systematic and high-flow sampling with small sample sizes to monitor temporal WQ changes, to reduce uncertainty in the RC parameters and to balance monitoring costs. The Chesapeake Bay WQ program employs a sampling program with 12 monthly routine samples plus eight storm samples at primary sites for nontidal WQ (EPA Chesapeake Bay program, 2017; Zhang & Hirsch, 2019). The USGS National Water Quality Network (NWQN) typically employs the program at six bimonthly fixed-interval samples plus 6–12 samples collected from months historically having increased loading (Lee et al., 2017). In contrast, monthly periodic samplings are typically employed in the WQ monitoring of public water areas in Japan. This combination of high-flow samples with periodic samples would be expected to reduce the bias of load estimates compared to the bias of the estimates based on periodic or random sampling by making the sampling probability approach PPS, though the degree of proxy to PPS, that is, the degree of bias of the estimates, is not controlled.

Considering the costs, labor, and importance of carrying out WQ monitoring in large rivers involving EWI sampling, depth-integrated sampling, or calibration of the representativeness of the samples at a fixed location over the river cross-section, we should conclude that the RCM using IS is currently not feasible at such monitoring sites. The RCM using IS has disadvantages for the application to monitoring sites in that it requires two-phase sampling and an inefficient sampling program limited to a single WQ parameter. This method also requires improvements for censored data sets and data sets with missing observations. We suggest the adoption of the SIR method to address these problems; the development of an improved method that applies the SIR method is a future topic of research. This means that the RCM using IS is still the LEM for small experimental catchments even when ΔL sampling is employed. The requirement of PPS samples for unbiased load estimation as described in this paper can help to determine

when high-flow samples should be collected in current WQ monitoring programs to reduce the load estimation bias. For example, discharge proportional sampling for every aliquot from the cumulative discharge would reduce the bias of load estimates for many WQ parameters compared with those based on periodic samples, although the degree of improvement is not controlled.

5. Conclusions

In this paper, we proposed the RCM using IS as an unbiased LEM and explained how to construct the CIs for river load estimates. We also provided a theoretical explanation of the unbiasedness of the proposed method and demonstrated the effectiveness of the proposed method based on highly frequent observation data. We revealed that the bias of the load estimates by the RCM does not originate from RC residual properties, including nonnormality and heteroscedasticity, sample size, watershed size, or land use in the watershed but mainly from the adoption of a biased load estimator, especially from an improper sampling method. However, many hurdles remain in applying this method to WQ monitoring sites, as this method does not have feasible solutions to missing and under-detection-limit observations and has an inefficient sampling strategy that is limited to the single WQ parameter. Although employing this method in actual monitoring sites in large rivers is currently difficult, the theoretical explanation of the unbiasedness of this method reveals what should be addressed for WQ sampling to ensure unbiased load estimation. The sampling probability should follow the PPS for this purpose. The feasibility of the proposed method is still limited to small research catchments even though ΔL sampling using an automatic sampler is adopted.

Conflict of Interest

The authors declare that they have no conflicts of interest.

Data Availability Statement

The discharge and concentration data used in this paper are available on Zenodo at the following address: https://doi.org/10.5281/zenodo.4485600. The US data used in this paper were originally downloaded from Heidelberg University's National Center for Water Quality Research site (https://ncwqr.org/monitoring/data) or the USGS National Water Information System (http://waterdata.usgs.gov/nwis/ or https://doi.org/10.5066/F7P55KJN).

Acknowledgments
First, the authors acknowledge the contributions from Eriko Yamamoto, Tomoya Ohira, Kazuya Tsuruga, Ryousuke Yoshimura, and Ryo Nishii to the development and improvement of the on-site monitoring system. The authors also thank the Kurahasi Giken Corp. for assembling the on-site monitoring system. The authors further acknowledge the contributions from Shuhei Kurihara to the evaluation of various LEMs and from Yuka Kuribayashi in the development and evaluation of the nonparametric interval estimation method for river loads. This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Numbers JP15K07646, JP20580263, and JP17780185.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/tac.1974.1100705

Appling, A. P., Leon, M. C., & McDowell, W. H. (2015). Reducing bias and quantifying uncertainty in watershed flux estimates: The R package loadflex. *Ecosphere*, *6*(12), 1–25. https://doi.org/10.1890/ES14-00517.1

Aulenbach, B. T. (2013). Improving regression-model-based streamwater constituent load estimates derived from serially correlated data. *Journal of Hydrology*, *503*, 55–66. https://doi.org/10.1016/j.jhydrol.2013.09.001

Aulenbach, B. T., Burns, D. A., Shanley, J. B., Yanai, R. D., Bae, K., Wild, A. D., et al. (2016). Approaches to stream solute load estimation for solutes with varying dynamics from five diverse small watersheds. *Ecosphere*, *7*(6), e01298. https://doi.org/10.1002/ecs2.1298

Aulenbach, B. T., & Hooper, R. P. (2006). The composite method: An improved method for stream-water solute load estimation. *Hydrological Processes*, *20*(14), 3029–3047. https://doi.org/10.1002/hyp.6147

Birgand, F., Faucheux, C., Gruau, G., Augeard, B., Moatar, B., & Bordenave, P. (2010). Uncertainties in assessing annual nitrate loads and concentration indicators: Part 1. Impact of sampling frequency and load estimation algorithms. *Transactions of the ASABE*, *53*(2), 437–446. https://doi.org/10.13031/2013.29584

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, *47*(5), 1287–1294. https://doi.org/10.2307/1911963

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). John Wiley & Sons.

Cochrane, D., & Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association*, *44*(245), 32–61. https://doi.org/10.1080/01621459.1949.10483290

Cohn, T. A. (1995). Recent advances in statistical methods for the estimation of sediment and nutrient transport in rivers. *Reviews of Geophysics*, *33*(S2), 1117–1123. https://doi.org/10.1029/95RG00292

Cohn, T. A. (2005). Estimating contaminant loads in rivers: An application of adjusted maximum likelihood to type 1 censored data. *Water Resources Research*, *41*(7). https://doi.org/10.1029/2004wr003833

Cohn, T. A., Caulder, D. L., Gilroy, E. J., Zynjuk, L. D., & Summers, R. M. (1992). The validity of a simple statistical model for estimating fluvial constituent loads: An empirical study involving nutrient loads entering Chesapeake Bay. *Water Resources Research*, *28*(9), 2353–2363. https://doi.org/10.1029/92wr01008

Cohn, T. A., Delong, L. L., Gilroy, E. J., Hirsch, R. M., & Wells, D. K. (1989). Estimating constituent loads. *Water Resources Research*, *25*(5), 937–942. https://doi.org/10.1029/WR025i005p00937

Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, *78*(383), 605–610. https://doi.org/10.1080/01621459.1983.10478017

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

Environmental Protection Agency (EPA) Chesapeake Bay Program. (2017). *Methods and quality assurance for Chesapeake Bay water quality monitoring programs, CBP/TRS-319-17*. Annapolis, MD: EPA. Retrieved from https://www.chesapeakebay.net/documents/CBPMethodsManualMay2017.pdf

Ferguson, R. I. (1986). River loads underestimated by rating curves. *Water Resources Research*, *22*(1), 74–76. https://doi.org/10.1029/WR022i001p00074

Gilroy, E. J., Hirsch, R. M., & Cohn, T. A. (1990). Mean square error of regression-based constituent transport estimates. *Water Resources Research*, *26*(9), 2069–2077. https://doi.org/10.1029/WR026i009p02069

Gulati, S., Stubblefield, A. A., Hanlon, J. S., Spier, C. L., & Stringfellow, W. T. (2014). Use of continuous and grab sample data for calculating total maximum daily load (TMDL) in agricultural watersheds. *Chemosphere*, *99*, 81–88. https://doi.org/10.1016/j.chemosphere.2013.10.026

Hammersley, J. M., & Handscomb, D. C. (1964). *Monte Carlo methods*. London: Methuen.

Hansen, M., & Hurwitz, W. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, *14*(4), 333–362. https://www.jstor.org/stable/2235923

Harmel, R. D., Cooper, R. J., Slade, R. M., Haney, R. L., & Arnold, J. G. (2006). Cumulative uncertainty in measured streamflow and water quality data for small watersheds. *Transactions of the ASABE*, *49*(3), 689–701. https://doi.org/10.13031/2013.20488

Harmel, R. D., Smith, D. R., King, K. W., & Slade, R. M. (2009). Estimating storm discharge and water quality data uncertainty: A software tool for monitoring and modeling applications. *Environmental Modelling & Software*, *24*(7), 832–842. https://doi.org/10.1016/j.envsoft.2008.12.006

Heidelberg University. (2019). *Heidelberg University National Center for Water Quality Research Tributary data download [data file]*. Retrieved from https://ncwqr.org/monitoring/data

Hirsch, R. M. (2014). Large biases in regression-based constituent flux estimates: Causes and diagnostic tools. *Journal of the American Water Resources Association*, *50*(6), 1401–1424. https://doi.org/10.1111/jawr.12195

Hollaway, M. J., Beven, K. J., Benskin, C. M. W. H., Collins, A. L., Evans, R., Falloon, P. D., et al. (2018). A method for uncertainty constraint of catchment discharge and phosphorus load estimates. *Hydrological Processes*, *32*(17), 2779–2787. https://doi.org/10.1002/hyp.13217

Horowitz, A. J. (2013). A review of selected inorganic surface water quality-monitoring practices: Are we really measuring what we think, and if so, are we doing it right? *Environmental Science & Technology*, *47*(6), 2471–2486. https://doi.org/10.1021/es304058q

International Reference Group on Great Lakes Pollution from Land Use Activities & Whitt, D. M. (1977). *Quality control handbook for pilot watershed studies*. International Joint Commission (IJC) Digital Archive. Retrieved from http://scholar.uwindsor.ca/ijcarchive/111

Johnes, P. J. (2007). Uncertainties in annual riverine phosphorus load estimation: Impact of load estimation methodology, sampling frequency, baseflow index and catchment population density. *Journal of Hydrology*, *332*(1–2), 241–258. https://doi.org/10.1016/j.jhydrol.2006.07.006

Keener, R. W. (2010). *Theoretical statistics topics for a core course*. New York, NY: Springer. Retrieved from https://doi.org/10.1007/978-0-387-93839-4

Koch, R. W., & Smillie, G. M. (1986). Bias in hydrologic prediction using log-transformed regression models. *Journal of the American Water Resources Association*, *22*(5), 717–723. https://doi.org/10.1111/j.1752-1688.1986.tb00744.x

Krishnamoorthy, K., & Mathew, T. (2003). Inferences on the means of lognormal distributions using generalized *p*-values and generalized confidence intervals. *Journal of Statistical Planning and Inference*, *115*(1), 103–121. https://doi.org/10.1016/s0378-3758(02)00153-2

Kronvang, B., & Bruhn, A. J. (1996). Choice of sampling strategy and estimation method for calculating nitrogen and phosphorus transport in small lowland streams. *Hydrological Processes*, *10*(11), 1483–1501. https://doi.org/10.1002/(sici)1099-1085(199611)10:11<1483::aid-hyp386>3.0.co;2-y

Kuhnert, P. M., Henderson, B. L., Lewis, S. E., Bainbridge, Z. T., Wilkinson, S. N., & Brodie, J. E. (2012). Quantifying total suspended sediment export from the Burdekin River catchment using the loads regression estimator tool. *Water Resources Research*, *48*(4), W04533. https://doi.org/10.1029/2011wr011080

Land, C. E. (1975). Tables of confidence limits for linear functions of the normal mean and variance. In Institute of Mathematical Statistics (Ed.), Selected tables in mathematical statistics (Vol. 3, pp. 385–419). Providence, RI: American Mathematical Society.

Lee, C. J., Hirsch, R. M., & Crawford, C. G. (2019). *An evaluation of methods for computing annual water-quality loads: U.S. Geological Survey Scientific Investigations Report 2019-5084*. Reston, VA: U.S. Geological Survey. https://doi.org/10.3133/sir20195084

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

Lee, C. J., Murphy, J. C., Crawford, C. G., & Deacon, J. R. (2017). *Methods for computing water-quality loads at sites: U.S. Geological Survey National Water Quality Network*. Open-File Report (Version 1.1: January 2020; Version 1.2: October 2017 ed.). Reston, VA: U.S. Geological Survey. Retrieved from https://doi.org/10.3133/ofr20171120

Lessels, J. S., & Bishop, T. F. A. (2013). Estimating water quality using linear mixed models with stream discharge and turbidity. *Journal of Hydrology*, *498*(Supplement C), 13–22. https://doi.org/10.1016/j.jhydrol.2013.06.006

Lessels, J. S., & Bishop, T. F. A. (2020). A post-event stratified random sampling scheme for monitoring event-based water quality using an automatic sampler. *Journal of Hydrology*, *580*, 123393. https://doi.org/10.1016/j.jhydrol.2018.12.063

Littlewood, I. G. (1992). *Estimating contaminant loads in rivers: A review* (Report No. 117). Wallingford OX: Institute of Hydrology.

Lloyd, C. E. M., Freer, J. E., Johnes, P. J., Coxon, G., & Collins, A. L. (2016). Discharge and nutrient uncertainty: Implications for nutrient flux estimation in small streams. *Hydrological Processes*, *30*(1), 135–152. https://doi.org/10.1002/hyp.10574

Mailhot, A., Rousseau, A. N., Talbot, G., Gagnon, P., & Quilbé, R. (2008). A framework to estimate sediment loads using distributions with covariates: Beaurivage River watershed (Québec, Canada). *Hydrological Processes*, *22*(26), 4971–4985. https://doi.org/10.1002/hyp.7103

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *8*(1), 3–30. https://doi.org/10.1145/272991.272995

McMillan, H., Krueger, T., & Freer, J. (2012). Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. *Hydrological Processes*, *26*(26), 4078–4111. https://doi.org/10.1002/hyp.9384

Moatar, F., & Meybeck, M. (2005). Compared performances of different algorithms for estimating annual nutrient loads discharged by the eutrophic River Loire. *Hydrological Processes*, *19*(2), 429–444. https://doi.org/10.1002/hyp.5541

Moatar, F., Person, G., Meybeck, M., Coynel, A., Etcheber, H., & Crouzet, P. (2006). The influence of contrasting suspended particulate matter transport regimes on the bias and precision of flux estimates. *Science of The Total Environment*, *370*(2–3), 515–531. https://doi.org/10.1016/j.scitotenv.2006.07.029

Owen, A. B. (1988). *Small sample central confidence intervals for the mean*. Department of Statistics, Technical Report (Vol. 302). Stanford, CA: Stanford University. Retrieved from https://statistics.stanford.edu/sites/default/files/EFS%20NSF%20302.pdf

Pagendam, D. E., Kuhnert, P. M., Leeds, W. B., Wikle, C. K., Bartley, R., & Peterson, E. E. (2014). Assimilating catchment processes with monitoring data to estimate sediment loads to the Great Barrier Reef. *Environmetrics*, *25*(4), 214–229. https://doi.org/10.1002/env.2255

Phillips, J. M., Webb, B. W., Walling, D. E., & Leeks, G. J. L. (1999). Estimating the suspended sediment loads of rivers in the LOIS study area using infrequent samples. *Hydrological Processes*, *13*(7), 1035–1050. https://doi.org/10.1002/(sici)1099-1085(199905)13:7<1035::aid-hyp788>3.0.co;2-k

Preston, S. D., Bierman, V. J., & Silliman, S. E. (1989). An evaluation of methods for the estimation of tributary mass loads. *Water Resources Research*, *25*(6), 1379–1389. https://doi.org/10.1029/WR025i006p01379

Raj, D. (1954). On sampling with probabilities proportionate to size. *Ganita*, *5*, 175–182.

Richards, R. P., & Holloway, J. (1987). Monte Carlo studies of sampling strategies for estimating tributary loads. *Water Resources Research*, *23*(10), 1939–1948. https://doi.org/10.1029/WR023i010p01939

Robert, C., & Casella, G. (2010). *Introducing Monte Carlo methods with R*. New York, NY: Springer. https://doi.org/10.1007/978-1-4419-1576-4

Rode, M., & Suhr, U. (2007). Uncertainties in selected river water quality data. *Hydrology and Earth System Sciences*, *11*(2), 863–874. https://doi.org/10.5194/hess-11-863-2007

Rubin, D. B. (1987). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, *82*(398), 543–546. https://doi.org/10.2307/2289460

Runkel, R. L., Crawford, C. G., & Cohn, T. A. (2004). *Load Estimator (LOADEST). A FORTRAN program for estimating constituent loads in streams and rivers*. Techniques and methods (Book 4, Chapter A5). Reston, VA: U.S. Geological Survey.

Rustomji, P., & Wilkinson, S. N. (2008). Applying bootstrap resampling to quantify uncertainty in fluvial suspended sediment loads estimated using rating curves. *Water Resources Research*, *44*(9), W09435. https://doi.org/10.1029/2007WR006088

Schleppi, P., Waldner, P. A., & Fritschi, B. (2006). Accuracy and precision of different sampling strategies and flux integration methods for runoff water: Comparisons based on measurements of the electrical conductivity. *Hydrological Processes*, *20*(2), 395–410. https://doi.org/10.1002/hyp.6057

Singh, A. K., Singh, A., & Engelhardt, M. (1997). *The lognormal distribution in environmental applications* (EPA/600/R-97/006). Washington, DC: U.S. EPA.

Slaets, J. I. F., Piepho, H.-P., Schmitter, P., Hilger, T., & Cadisch, G. (2017). Quantifying uncertainty on sediment loads using bootstrap confidence intervals. *Hydrology and Earth System Sciences*, *21*(1), 571–588. https://doi.org/10.5194/hess-21-571-2017

Slaets, J. I. F., Schmitter, P., Hilger, T., Lamers, M., Piepho, H.-P., Vien, T. D., & Cadisch, G. (2014). A turbidity-based method to continuously monitor sediment, carbon and nitrogen flows in mountainous watersheds. *Journal of Hydrology*, *513*, 45–57. https://doi.org/10.1016/j.jhydrol.2014.03.034

Stenback, G. A., Crumpton, W. G., Schilling, K. E., & Helmers, M. J. (2011). Rating curve estimation of nutrient loads in Iowa rivers. *Journal of Hydrology*, *396*(1–2), 158–169. https://doi.org/10.1016/j.jhydrol.2010.11.006

Tada, A., Tanakamaru, H., & Hata, T. (2006). Long-term and high temporal resolution in-situ monitoring of potassium, sodium, and chloride in a small forested catchment using flow injection potentiometry. *Journal of Japan Society of Hydrology & Water Resources*, *19*(6), 445–457. https://doi.org/10.3178/jjshwr.19.445

Tanner, M. A. (2006). *Tools for statistical inference* (3rd ed.). New York, NY: Springer. Retrieved from https://doi.org/10.1007/978-1-4684-0192-9

Thomas, R. B. (1985). Estimating total suspended sediment yield with probability sampling. *Water Resources Research*, *21*(9), 1381–1388. https://doi.org/10.1029/WR021i009p01381

Thomas, R. B. (1988a). *Measuring sediment yields of storms using PSALT*. Sediment budgets (Vol. 174, pp. 315–323). IAHS Publication.

Thomas, R. B. (1988b). Monitoring baseline suspended sediment in forested basins: The effects of sampling on suspended sediment rating curves. *Hydrological Sciences Journal*, *33*(5), 499–514. https://doi.org/10.1080/02626668809491277

Thomas, R. B. (1989). *Piecewise SALT sampling for estimating suspended sediment yields* (General Technical Report, PSW-114). Berkeley, CA: Pacific Southwest Forest and Range Experiment Station, U.S. Forest Service.

Thomas, R. B., & Lewis, J. (1993). A comparison of selection at list time and time-stratified sampling for estimating suspended sediment loads. *Water Resources Research*, *29*(4), 1247–1256. https://doi.org/10.1029/92wr02711

Thomas, R. B., & Lewis, J. (1995). An evaluation of flow-stratified sampling for estimating suspended sediment loads. *Journal of Hydrology*, *170*(1–4), 27–45. https://doi.org/10.1016/0022-1694(95)02699-p

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, *26*(1), 24–36. https://doi.org/10.2307/1907382

U.S. EPA (2002). *Calculating upper confidence limits for exposure point concentrations at hazardous waste sites* (OSWER 9285.6-10). Washington, DC: Office of Emergency and Remedial Response, U.S. EPA.

U.S. Geological Survey. (2016). *National water information system data available on the World Wide Web (USGS water data for the nation) [Data file]*. Retrieved from http://waterdata.usgs.gov/nwis/, https://doi.org/10.5066/F7P55KJN

Valentine, H. T., Tritton, L. M., & Furnival, G. M. (1984). Subsampling trees for biomass, volume, or mineral-content. *Forest Science*, *30*(3), 673–681.

Verhoff, F. H., Yaksich, S. M., & Melfi, D. A. (1980). River nutrient and chemical-transport estimation. *Journal of the Environmental Engineering Division*, *106*(3), 591–608.

Verma, S., Markus, M., & Cooke, R. A. (2012). Development of error correction techniques for nitrate-N load estimation methods. *Journal of Hydrology*, *432-433*, 12–25. https://doi.org/10.1016/j.jhydrol.2012.02.011

Vigiak, O., & Bende-Michl, U. (2013). Estimating bootstrap and Bayesian prediction intervals for constituent load rating curves. *Water Resources Research*, *49*(12), 8565–8578. https://doi.org/10.1002/2013wr013559

Walling, D. E., & Webb, B. W. (1981). *The reliability of suspended sediment load data*. In Erosion and sediment transport measurement (Vol. 133, pp. 177–194). IAHS Publication.

Wang, Y. G., Kuhnert, P., & Henderson, B. (2011). Load estimation with uncertainties from opportunistic sampling data? A semiparametric approach. *Journal of Hydrology*, *396*(1–2), 148–157. https://doi.org/10.1016/j.jhydrol.2010.11.003

Webb, B. W., Phillips, J. M., Walling, D. E., Littlewood, I. G., Watts, C. D., & Leeks, G. J. L. (1997). Load estimation methodologies for British rivers and their relevance to the LOIS RACS(R) programme. *Science of The Total Environment*, *194–195*, 379–389. https://doi.org/10.1016/s0048-9697(96)05377-6

Weerahandi, S. (1995). *Exact statistical methods for data analysis*. New York, NY: Springer.

Yanai, R. D., Tokuchi, N., Campbell, J. L., Green, M. B., Matsuzaki, E., Laseter, S. N., et al. (2015). Sources of uncertainty in estimating stream solute export from headwater catchments at three sites. *Hydrological Processes*, *29*(7), 1793–1805. https://doi.org/10.1002/Hyp.10265

Zamyadi, A., Gallichand, J., & Duchemin, M. (2007). Comparison of methods for estimating sediment and nitrogen loads from a small agricultural watershed. *Canadian Biosystems Engineering*, *49*(1), 1.27–1.36.

Zhang, Q., & Hirsch, R. M. (2019). River water-quality concentration and flux estimation can be improved by accounting for serial correlation through an autoregressive model. *Water Resources Research*, *55*(11), 9705–9723. https://doi.org/10.1029/2019wr025338