Unbiased Estimates and Confidence Intervals of Riverine Loads for Low-Frequency Water Quality Monitoring Strategies

Tada, Akio

Tanakamaru, Haruya

(Citation)
Water Resources Research,58(5):e2022WR031941

(Issue Date)
2022-05

(Resource Type)
journal article

(Version)
Version of Record

Water Resources Research

RESEARCH ARTICLE

10.1029/2022WR031941

Key Points:

- The Horvitz-Thompson estimator can provide unbiased load estimates based on various sampling strategies
- We propose an efficient interval estimation method for riverine loads based on the Horvitz-Thompson estimator
- A procedure for load estimation with missing observations is proposed

Supporting Information:

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

Correspondence to:

A. Tada,
atada@kobe-u.ac.jp

Citation:

Tada, A., & Tanakamaru, H. (2022). Unbiased estimates and confidence intervals of riverine loads for low-frequency water quality monitoring strategies. *Water Resources Research*, *58*, e2022WR031941. https://doi.org/10.1029/2022WR031941

Received 5 JAN 2022
Accepted 18 APR 2022

Author Contributions:

Conceptualization: Akio Tada, Haruya Tanakamaru
Funding acquisition: Akio Tada, Haruya Tanakamaru
Investigation: Akio Tada
Methodology: Akio Tada, Haruya Tanakamaru
Project Administration: Akio Tada
Resources: Haruya Tanakamaru
Software: Akio Tada
Supervision: Haruya Tanakamaru
Validation: Akio Tada
Writing – original draft: Akio Tada

Unbiased Estimates and Confidence Intervals of Riverine Loads for Low-Frequency Water Quality Monitoring Strategies

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

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

Abstract Accurate estimates of riverine constituent loads are essential for water pollution control and erosion control in watersheds, but an accurate estimation method for commonly employed water quality monitoring strategies, such as calendar-based sampling with high-flow sampling, has not been established. Therefore, we propose methods for both unbiased load estimation and confidence interval construction for annual riverine loads using the Horvitz-Thompson estimator based on limited samples (12 to 20 per year) collected with commonly used water quality monitoring strategies. In addition, we propose an uncertainty reduction method to calculate efficient confidence intervals using probability resampling based on a simple rating curve due to the small sample sizes. The effectiveness of the proposed method is verified by load estimates based on 150 annual data sets of daily pairs of discharge and concentration observations of six water quality parameters from different watersheds and years. The tested sampling strategies include random sampling over time as a proxy for calendar-based sampling, random sampling with high-flow sampling, and flow-proportional sampling. The results show that the proposed methods provide unbiased load estimates for all the data sets and appropriate confidence intervals. The uncertainty reduction in the confidence interval widths obtained with probability resampling is effective for calendar-based sampling. We also explain the importance of defining proper sampling probabilities for load estimation based on samples with missing observations. The proposed method can be used to evaluate and improve existing water quality monitoring strategies for more efficient load estimation and the retrieval of accurate load estimates from existing data sets.

1. Introduction

Eutrophication in semiclosed water bodies, such as lakes and estuaries, due to excessive nutrient loads has severely damaged the ecosystem via water pollution. The establishment of pollution prevention measures in ecosystems requires accurate estimates of the total mass of pollutants flowing into a water body during a certain period. Because rivers are the main transport paths of pollutants to a receiving water body, accurate load estimation methods for riverine constituent loads have been studied over the past four decades. These studies assumed that river discharge was observed at a high frequency, such as daily, hourly, or subhourly, but water quality (WQ) is generally observed at a low frequency, such as weekly or monthly, due to the high cost and labor requirements for sampling and analysis. This mismatch among observation frequencies can result in highly uncertain load estimates.

The low-frequency sampling methods used for load estimation include autosampling, stratified sampling (Thomas & Lewis, 1993, 1995), composite sampling (Horowitz, 2010), probability sampling (Thomas, 1985, 1988, 1989), and grab sampling. Lessels and Bishop (2020) proposed a posterior stratified sampling method, but a feasible strategy for strata and sample-size allocation for precise load estimation has not been established. Flow-proportional composite sampling is expected to provide a good proxy for true loads, but it requires continuous or high-frequency composite WQ samples to obtain good approximations. Probability sampling is rarely employed in practice because random sampling at monitoring sites was recommended by Thomas (1985). A highly frequent bankside autoanalyzer can provide a reliable value of the true load, as previously reported (Cassidy & Jordan, 2011), but the broad application of such devices at monitoring sites is challenging due to the high initial and maintenance costs. Miller et al. (2017) used high-frequency in-stream nitrate measurements obtained with a nitrate sensor to estimate nonpoint nitrate loads to streams, although the sensor would restrict measurable WQ parameters. The above sampling methods, except grab sampling, have a common limitation: these sampling methods are often infeasible for large rivers, where the distribution of discharge and WQ concentrations are not uniform over the river cross-section. Grab sampling is a low-cost sampling method and can provide representative WQ samples over a cross-section using the equal width increment (EWI) sampling approach (Guy & Norman, 1970); however, grab sampling cannot provide high-frequency WQ samples. As a result, monthly calendar-based sampling was employed by the EU (Jung et al., 2010), Japan (Luo et al., 2013), and the U.S. from the 1970s to the 1990s (Horowitz, 2015).

In studies focused on developing a less biased load estimation method using low-frequency WQ observations, the unbiasedness (lack of bias) of load estimates has been generally ignored, except in a few cases. The use of a biased estimator may be justified if the estimator can be easily applied or the feasibility of sampling is the most important factor due to costs and monitoring labor requirements. Even in such cases, the substantial estimation error associated with a biased estimator can be problematic (e.g., Cohn et al., 1989). Raj (1968) described the conditions for justifying the use of a biased estimator: the coverage of confidence intervals (CIs) with a biased estimator should be almost the same as that with an unbiased estimator. Because the design of such a biased estimator is impossible without knowing the distribution of population elements, an unbiased estimator should be used in load estimation. Studies of unbiased load estimation have been based on either stratified sampling (Thomas & Lewis, 1993, 1995; Verhoff et al., 1980) that used the weighted average of stratum means as a load estimate based on simple random sampling (SRS), or probability sampling methods (Thomas, 1985, 1988, 1989). In other studies, load calculation methods (LCMs) have been independently combined with sampling strategies. By ignoring the logical connection between an LCM and the strategy (Thomas & Lewis, 1995), the unbiasedness of estimates has been impaired. As a result, the bias of load estimates, or the difference between the expectation of load estimates and the true loads, could not be controlled; therefore, bias was empirically investigated for various watersheds, WQ parameters, and sampling strategies (e.g., Lee et al., 2016; Preston et al., 1989; Walling & Webb, 1981) instead of maintaining a zero bias. However, in such an empirical approach, there is no guarantee that a load estimation method that is efficient (precise) and less-biased based on samples collected with a particular sampling strategy (sampling procedure and sample size) in a certain watershed can perform well for other samples obtained with a different sampling strategy in a different watershed. Because biased estimates could lead to inappropriate water pollution control measures and best management practices employed for nonpoint pollutant load reduction, unbiased load estimates are required for effective water pollution control.

Thomas (1985) proposed the selection at list time (SALT) method, an efficient and unbiased load estimation method, based on the sampling probability proportional to size (PPS) concept. The SALT method is a realization of PPS sampling at WQ monitoring sites using a cumulative-total (e.g., Arnab, 2017) or inverse function method. Tada and Tanakamaru (2021) explained that PPS sampling is a type of discretized importance sampling (IS, e.g., Hammersley & Handscomb [1964]) and proposed the rating curve method (RCM) using IS as a practical method for obtaining unbiased load estimates and the corresponding CIs. However, PPS sampling is not feasible at WQ monitoring sites because it requires real-time observation of the auxiliary variables related to loading rates and a programmable autosampler. Therefore, an unbiased load estimation method should be investigated based on WQ samples collected with feasible and commonly used low frequency WQ monitoring strategies, such as calendar-based sampling. The feasibility of a sampling strategy in terms of labor and cost is critical in WQ monitoring programs.

Based on the above discussion, this study has three objectives. The first objective is to verify the effectiveness of the Horvitz-Thompson (HT) estimator (Arnab, 2017; Corchran, 1977; Lohr, 2021) in annual river load estimation based on a limited (small-size) samples collected with a feasible and practical sampling method; although the HT estimator is a popular estimator, it has rarely been applied for load estimation. In this paper, the term sample refers to a subset of the population, and a member of a sample or the population is referred to as an element. The second objective is to propose a method for configuring the CIs of load estimates based on the HT estimator. The final objective is to develop a method for precisely determining the CIs of load estimates for a sample. To achieve the last objective, a probability resampling method that resamples elements with sampling probabilities that are proportional to the magnitude of estimated river loading rate (river load in a short time interval) from a non-PPS sample is proposed. As is common in other load estimation studies, this study assumes that discharge data are observed almost continuously, such as daily or subdaily, but WQ data are observed at a low frequency, such as weekly or monthly. To validate the effectiveness of the methods proposed, 150 annual data sets of daily pairs of discharge and WQ observations with less than 15 missing observations per year are used. Evaluated data sets for six WQ parameters, suspended solids (SS), total phosphorous (TP), soluble reactive phosphorous (SRP), nitrite and nitrate (NO23), total Kjeldahl nitrogen (TKN), and chloride (Cl), are considered. The data are obtained from nine watersheds in different years (25 combinations of watersheds and years) by Heidelberg University's National Center for Water Quality Research (NCWQR, Heidelberg University [2020]). Both the validation of the unbiasedness of load estimates and the appropriateness of 95% central CIs are based on small size samples with 12–20 elements extracted from the data sets by feasible sampling strategies. The tested sampling strategies include monthly random (RDM) sampling with 12 elements as a proxy for monthly calendar-based sampling, monthly random sampling plus high-flow random (RHF) sampling with 20 elements, and flow-proportional (FP) sampling with 12 elements. Validation was based on the load estimates of 20,000 sets of resampled samples from each data set. Through the validation of the proposed methods, we confirm that these methods can be used for improving existing WQ monitoring strategies and retrieving unbiased load estimates from existing observed data sets for improved water pollution control.

It should be noted that the uncertainties discussed in this paper are associated with the LCM and sampling strategy used; we do not discuss the uncertainty from errors in discharge observations (Lloyd et al., 2016), errors in WQ analysis (McMillan et al., 2012), or errors in the representativeness of WQ samples over a river cross-section (Horowitz, 2013; Rode & Suhr, 2007). We also do not discuss the errors from diurnal WQ variations (Jones et al., 2012) or errors from the mismatch between the temporal units of discharge data and WQ samples (Lee et al., 2017) because the evaluations in this paper are based on daily paired WQ and discharge data.

2. Riverine Load Estimation With the HT Estimator

2.1. Unbiased Load Estimation With the HT Estimator

First, we should clarify the definition of unbiasedness. When we estimate an annual riverine load L using a model such as a rating curve (RC), the unbiasedness of the load estimates $\hat{L}$ is defined as follows (e.g., Keener et al., 2010):

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

where L_{TRUE} is the true load value and $\mathrm{E}_{\theta}\,[\cdot]$ denotes the expectation of $\cdot$ for model parameter θ. Bias is defined as the difference between the expectation of $\hat{L}$ and L_{TRUE}. Commonly, accuracy relates to bias, and precision refers to the variance (the second moment around the expectation of $\hat{L}$). Because there is no guarantee that the value of θ is known for a population, unbiasedness requires zero bias for any estimated θ value. Because we cannot control the magnitude of bias, we can only use an unbiased estimator. An unbiased estimator satisfies the conditions in Equation 1 for any sample of any size, from any watershed, and for any WQ parameter.

Common misconceptions about the unbiasedness of load estimates are mainly related to the biased but consistent estimates obtained with interpolation (e.g., Walling & Webb, 1981) or composite methods (Aulenbach, 2013; Aulenbach et al., 2016; Aulenbach & Hooper, 2006). Although the bias of a consistent estimator approaches zero as the sample size increases, a consistent but biased estimator has nonzero bias when the sample size is small. Since a consistent estimator requires many samples to keep the bias zero, we should not employ a consistent but biased estimator for small sample sizes. It should also be noted that an increase in sample size reduces not bias but variance unless an estimator is consistent. Another common misconception concerns the bias correction factor (BCF) used in the RCM. BCFs are used in the RCM to correct log transformation bias. Representative RCM estimators with BCFs include the quasimaximum likelihood estimator (QMLE, Cohn et al., 1989; Ferguson, 1986a; Thomas, 1985), smearing estimator (Duan, 1983; Koch & Smillie, 1986a; Thomas, 1985), and minimum variance unbiased estimator (MVUE, Cohn et al. [1989]). These BCFs correct only the log transformation bias under the assumption that log residuals of an RC follow a normal distribution. In fact, these BCFs cannot correct the bias associated with the sampling strategy or the nonnormality of the residual distribution. This limitation of QMLE was discussed in 1986 by Ferguson (1986b) and Koch and Smillie (1986b). This discussion of QMLE also holds for the smearing estimator and MVUE. To keep load estimates obtained with the RCM unbiased, samples acquired with sampling probabilities proportional to the expected loading rates by an RC must be used (Tada & Tanakamaru, 2021; Thomas, 1985).

Here, we consider the relationship between sampling strategies and unbiased load estimates. Suppose that we calculate the river load L_{TRUE} from time t_1 to t_2. In this calculation, considering the sampling probability $p(t)$, we integrate the instantaneous loading rate $l(t)$ over $d\eta = p(t)dt$ instead of dt:

$$L_{TRUE} = \int_{t_1}^{t_2} l(t)dt = \int_{t_1}^{t_2} \frac{l(t)}{p(t)} p(t)dt = \int_0^1 \frac{l(\eta)}{p(\eta)} d\eta \tag{2}$$

The period $[t_1, t_2]$ is divided into N constant time intervals (Δts) between observations, and Δt is short enough to assume that $l(t)$ is constant. When n sample elements are collected with the sampling probability p_j for the jth element ($j = 1$ to N), the HT estimator L_{HT} (Horvitz & Thompson, 1952) for point estimation is defined as:

$$L_{HT} = \frac{1}{n} \sum_{i=1}^{n} \frac{l_i}{p_i} = \sum_{i=1}^{n} \frac{l_i}{\pi_i} \tag{3}$$

where l_i, p_i, and π_i are the loading rate, sampling probability, and inclusion probability of the ith sample element ($i = 1$ to n), respectively. The values of p_i and π_i must be greater than zero. Because the expectation of L_{HT} is L_{TRUE} (e.g., Gregoire & Valentine, 2007) from Equations 2 and 3, L_{HT} is an unbiased estimator. WQ sampling without duplicated sample elements with probability p_j can be regarded as unequal probability sampling without replacement (UPSWOR), and the HT estimator is an unbiased estimator for samples obtained with UPSWOR. In this case, π_j is the expected contribution of the jth sample to a sample of size n, and the sum of π_j for N population elements becomes n ($p_i = \pi_i/n$, Arnab [2017]; Lohr [2021]). When π_i is proportional to the magnitude of elements (inclusion probability proportional to size, IPPS), L_{HT} becomes optimally precise. In the case of unequal probability sampling with replacement (UPSWR) and p_j proportional to the magnitude of elements, Equation 3 is referred to as the Hansen-Hurwitz (HH) estimator (Hansen & Hurwitz, 1943); this estimator has been used as the LCM in the SALT and piecewise SALT methods for PPS sampling with duplicated sample elements (Thomas, 1985, 1988, 1989). In summary, after p_j for a sampling strategy is determined, unbiased load estimates can be calculated based on Equation 3. It should be noted that L_{HT} is no longer unbiased for the population total when $p_j = 0$ for some elements of the population (nonsampling error). Arabkhedri et al. (2010) used the modified HT estimator with adaptive cluster sampling for estimating the total suspended sediment load but reported considerable underestimation for small samples of less than 15% of the population. They attributed this underestimation to nonsampling error. The nonsampling error problem will be discussed in 5.2.

Now, we demonstrate two examples of the logical connection between the sampling strategy and an LCM for unbiased load estimation based on the HT estimator. Two sampling strategies, random sampling over time and FP sampling are used. For a sample with n elements collected through SRS over the period $[t_1, t_2]$, π_j is defined as n/N, and $p_j = 1/N$. In this case, L_{HT} becomes:

$$L_{HT} = \sum_{i=1}^{n} \frac{l_i}{\pi_i} = \frac{N}{n} \sum_{i=1}^{n} l_i = N\bar{l}_s \tag{4}$$

where $\bar{l}_s$ is the mean loading rate of a sample. In this case, the estimator is an average estimator. For a sample with n elements collected through FP sampling over the period $[t_1, t_2]$, $p_j = q_j/Q_T$, where q_j is the discharge for the jth element and Q_T is the total discharge of N elements in the population. In this case, L_{HT} becomes:

$$L_{HT} = \frac{1}{n} \sum_{i=1}^{n} \frac{l_i}{p_i} = \frac{Q_T}{n} \sum_{i=1}^{n} \frac{l_i}{q_i} = Q_T \frac{1}{n} \sum_{i=1}^{n} c_i = \bar{c}_s Q_T \tag{5}$$

where $\bar{c}_s$ is the mean concentration of a sample. This estimator is another average estimator. In FP sampling, L_{HT} in Equation 5 is the most precise when l_j is proportional to q_j but remains unbiased regardless of the relationship between l_j and q_j.

In addition, the load estimator, as a stratum-size weighted average of the stratum means with stratified sampling based on SRS (Thomas & Lewis, 1995; Verhoff et al., 1980), is also an HT estimator, although the ratio estimator (Beale, 1962) with stratified sampling employed in Great Lake studies (International Reference Group on Great Lakes Pollution from Land Use Activities & Whitt, 1977) was biased. The load estimates by the RCM become an HT estimator when p_j in UPSWOR is proportional to the estimated loading rates by an RC (Tada &

Tanakamaru, 2021). As shown here, for unbiased load estimation, the relationship between an LCM and sampling probability defined by the sampling strategy must satisfy Equation 3; otherwise, load estimates become biased.

2.2. Sampling Probabilities for RDM, RHF, and FP Sampling

The representative sampling strategies employed at WQ monitoring sites include monthly calendar-based sampling used in the National Stream Quality Accounting Network (NASQAN) program of the USGS from the 1970s to the 1990s (Horowitz, 2015). Because monthly calendar-based sampling involves the collection of more low-flow elements than high-flow elements, NASQAN revised its strategy so that sample elements could be collected in a wider range of hydrological regimes. The revised strategy adopted in 2007 involves collecting six bimonthly calendar-based samples and six samples that reflect various hydrological and seasonal changes (USGS, 2009). In the National Water Quality Network (NWQN) established in 2013, six bimonthly calendar-based samples and 6–12 samples from months historically characterized by increased loading are collected (Lee et al., 2017; USGS, 2021). Other representative strategies include a combination of calendar-based sampling and high-flow sampling. For example, the EU adopts fortnightly to monthly periodic sampling plus high-flow sampling during floods (Water Framework Directive, 2003), and the Chesapeake Bay Program (CBP) collects 12 monthly calendar-based samples and eight targeted storm samples (Chanat et al., 2016: EPA Chesapeake Bay Program, 2017). These combined strategies of periodical and high-flow sampling limit the number of sample elements to 12–20 annually due to high labor and costs in WQ monitoring. In addition, high-flow sampling requires the establishment of a threshold to distinguish between high flow and low flow. For example, Oelsner et al. (2017) used the 85th percentile of decadal daily discharge for each month as the threshold to investigate long-term WQ trends. Lee et al. (2019) employed the 80th percentile and Zhang and Hirsch (2019) used the 90th percentile of daily discharge in a year as thresholds in their high-flow sampling simulations.

In this study of load estimation with the HT estimator, three types of sampling are considered: RDM sampling, RHF sampling, and FP sampling (Figure 1). Whereas operating monthly calendar-based sampling at WQ monitoring sites would be neither monthly random sampling nor systematic sampling with a fixed interval in a strict sense, we employed RDM sampling (Figure 1b); in this approach, one sample element was collected by SRS in each month as a proxy for calendar-based sampling. Although systematic sampling is a type of equal probability sampling similar to SRS, it may produce biased load estimates when the population exhibits cyclical variation. Systematic sampling also tends to provide more precise estimates than SRS (Cochran, 1977; Zamyadi et al., 2007). In addition, samples collected by systematic sampling constitute subsets of the population that have a particular statistical characteristic when the population size is not large, such as 365 daily observations per year. Consequently, the bias of load estimates based on systematic samples obtained with the Monte Carlo method may not be zero. Considering the above factors, we employed random sampling instead of systematic sampling in load estimation. In a year with N days, the sampling probability p_{0jk} for the jth day ($j = 1$ to N) in the kth month ($k = 1$ to 12) is:

$$p_{0jk} = \frac{1}{n_1} \times \frac{n_m}{N_{mk}} \tag{6}$$

where n_m is the number of sample elements allocated for each month (in this study, $n_m = 1$), N_{mk} is the number of days in the kth month, and n_1 is the sample size in a year (in this study, $n_1 = n_m \times 12 = 12$). Since N_{mk} is different among months, RDM sampling in this study is UPSWOR.

RHF sampling (Figures 1a and 1c) is a combination of RDM sampling with high-flow sampling and mimics the strategy employed in the CBP. In this study, the threshold between high flow and low flow was set at the 90th percentile of discharge (q_{90}, with a non-exceedance probability of 0.90) in a year, as used by Zhang and Hirsch (2019). A high-flow sampling method was designed to collect eight samples randomly on days with discharge values that were larger than or equal to q_{90}. In this case, the sampling probability p_{0jk} becomes;

$$p_{0jk} = \frac{1}{n_1 + n_2} \times \frac{n_m}{N_{mk}} \quad (q_j < q_{90}) \tag{7a}$$

$$p_{0jk} = \frac{1}{n_1 + n_2} \times \left(\frac{n_m}{N_{mk}} + \frac{n_2}{N_h} \right) \quad (q_j \geq q_{90}) \tag{7b}$$

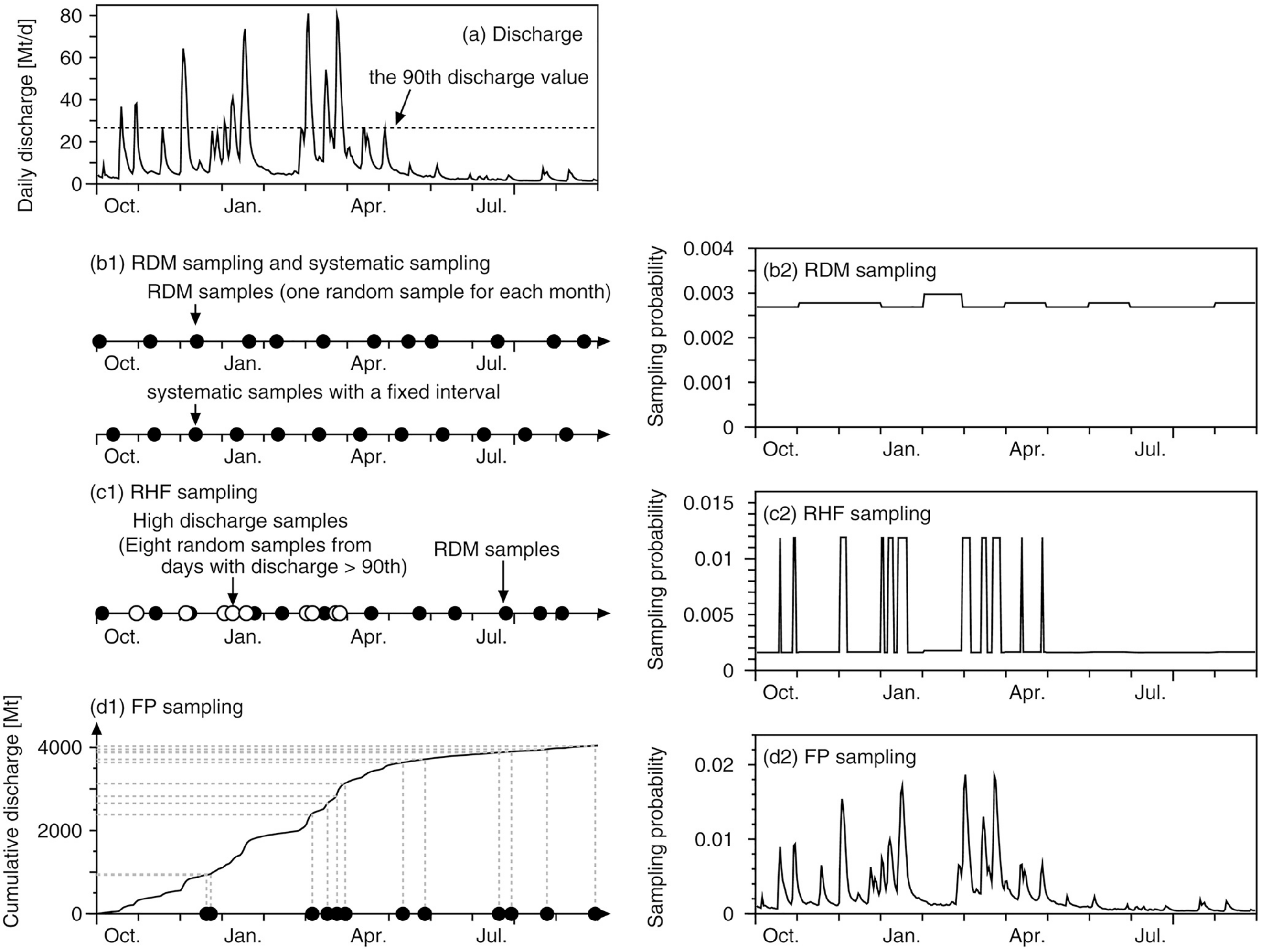


Figure 1. Schematic diagrams of monthly random sampling, high flow random, and flow proportional sampling and the corresponding sampling probabilities.

where n_2 is the number of high-flow sample elements (in this study, $n_2 = 8$) and N_h is the number of high-flow elements in the population. RHF sampling involves collecting 20 sample elements in a year. In our scenario of RHF sampling, RDM sampling was performed first, and then high-flow sampling was conducted while avoiding the duplication of elements. Because the value of q_{90} in a year is unknown in advance, the RHF sampling approach employed here is a retrospectively definable strategy. This strategy is also a form of UPSWOR.

FP sampling (Figure 1d) involved collecting 12 sample elements per year and was implemented based on the cumulative-total method in this study. In the cumulative-total method, CQ_j ($j = 1$ to N), the cumulative discharge on the jth day, was calculated by summing the daily discharge values from the first day of a water year to the jth day ($CQ_0 = 0$ and $CQ_N = Q_T$). Then, a random number rdn_i ($i = 1$ to 12) in the range of (0, Q_T] was generated, and the jth population element was obtained as the ith sample element when $CQ_{j\text{-}1} < rdn_i \leq CQ_j$. In this case, the sampling probability for the jth element p_{0j} is given as follows, as long as duplication in sample elements does not occur:

$$p_{0j} = \frac{q_j}{Q_T} \tag{8}$$

In FP sampling without replacement (UPSWOR), only one sample element is selected, even when multiple rdn_i values are included in ($CQ_{j\text{-}1}$, CQ_j]. As shown in Figure S1 in Supporting Information S1, p_{0j} is not linearly related to q_j, as in Equation 8, when $n > Q_T/q_{max}$, where q_{max} is the maximum daily discharge in a year. Notably, the sampling probability on a given day with a large discharge, such as q_{max}, tends to be associated with a p_{0j} value larger than the probability defined in Equation 8. In this case, p_{0j} must be determined by Monte Carlo simulation, as described later in Section 3.2.2.

Discretized systematic FP sampling using a fixed aliquot of discharge ΔQ would be more feasible at WQ monitoring sites than sampling with the cumulative-total method. However, regardless of which sample elements are collected by the cumulative-total method or systematic discretized sampling in FP sampling, we will not know Q_T in advance; we also cannot determine the sample size n in advance. In addition to RHF sampling, the FP sampling method employed here was a retrospectively definable strategy. When FP sampling is applied at an actual WQ monitoring site, the sample size must be adjusted based on the observed and planned Q_T values, as in the SALT method (Thomas, 1985). Additionally, when the daily discharge on a given day cannot be reported in time for sampling, sampling should be performed under a certain assumption that the daily discharge value on that day is the same as that on the prior day, for example, Although FP sampling is difficult to implement in practice at actual WQ monitoring sites, we employed FP sampling as an efficient WQ sampling method for load estimation (e.g., Schleppi et al., 2006).

2.3. Uncertainty Reduction Through Probability Resampling

The HT estimator achieves almost minimum variance when the inclusion probability is roughly proportional to the magnitude of the population elements, for example, to approximate loading rates $\hat{l}(t)$, because considerable variation in $l(t)$ can be expressed by the variation in its approximate value. Accordingly, when sample elements with inclusion probabilities proportional to $\hat{l}(t)$ can be resampled from a sample collected with non-IPPS samplings (a non-IPPS sample), such as RDM, RHF, or FP sampling, more precise estimates than the HT estimates calculated from all the elements of a non-IPPS sample can be derived. However, resampling cannot always reduce the variance in estimates based on resampled IPPS samples because resampling is accompanied by a reduction in sample size. Whether this probability resampling approach (IPPS resampling) can reduce the variance of HT estimates based on an original sample depends on both the positive effect of the improved proportionality of the inclusion probabilities of resampled elements to loading rates and the negative effect of sample size reduction.

Resampling with sampling probability p_{1i} ($i = 1$ to n) from a sample of size n collected at sampling probability p_{0i} can be performed by rejection sampling (RS, Bishop [2006]; Lahiri [1951]). The sampling probability p_{1i} should be proportional to the loading rate estimated by an RC to obtain efficient load estimates. In this study, we employed the simplest RC for resampling as:

$$\ln \hat{l}_j = \beta_0 + \beta_1 \ln q_j \tag{9}$$

where $\hat{l}_j$ and q_j are the estimated loading rate and discharge of the jth population element, respectively, and β_0 and β_1 are regression coefficients. Although more sophisticated RCs than that expressed in Equation 9, such as the LOADEST models (Runkel et al., 2004), WRTDS (Hirsch, 2014; Hirsch et al., 2010), or WRTDS-K (Lee et al., 2019; Zhang & Hirsch, 2019), are generally used to obtain load estimations with multiple explanatory variables for U.S. rivers, we use a simple RC. Notably, sophisticated RCs require large data sets that span multiple years, such as 60 sample elements per 10 years in load estimation for the WRTDS model (Hirsch & De Cicco, 2015). The small sample size targeted in this study, ranging from 12 to 20 sample elements for annual load estimation, would inflate the uncertainty of partial coefficients of multiple explanatory variables and, in turn, increase the uncertainty of load estimates by such RCs (Tada & Tanakamaru, 2021). Sophisticated RCs are superior to the simple RC in Equation 9 when data sets with longer periods, such as decadal data, are used in load estimation.

The resampling probability p_{1i} based on Equation 9 is defined as:

$$p_{1i} = \hat{l}_i \Big/ \sum_{j=1}^{N} \hat{l}_j = q_i^{\beta_1} \Big/ \sum_{j=1}^{N} q_j^{\beta_1} \tag{10}$$

Therefore, the weight of the ith sample element used for rejection sampling based on IPPS (IPPS RS) becomes:

$$w_i = \left(\frac{p_{1i}}{p_{0i}}\right) \Big/ \left(\frac{p_1}{p_0}\right)_{\max} \tag{11}$$

where $(p_1/p_0)_{\max}$ is the maximum value of (p_{1j}/p_{0j}) for the population (j = 1 to N). When $(p_1/p_0)_{\max}$ is defined as the maximum value of (p_{1i}/p_{0i}) for a sample (i = 1 to n), w_i becomes the weight for sampling/importance resampling (SIR, Rubin [1987]; Tanner [2006]). Rubin (1987) recommended a sample size of 20 times the size of a resampled sample, and the SIR method should not be applied to a sample with a small size. In RS, n uniform random numbers with a range of (0, 1], $rdn0_i$s (i = 1 to n), are generated, and each $rdn0_i$ is compared to w_i. The ith sample element is rejected (not resampled) when $rdn0_i$ is greater than w_i, and the ith element is accepted (resampled) when $rdn0_i$ is equal to or less than w_i. It should be noted that the sampling probability p_1 does not need to be proportional to the estimated loading rates when minimum variance is not essential, that is, any sampling probability is applicable to p_1 for unbiased estimation.

2.4. Construction of CIs for Load Estimates

Sampling textbooks (e.g., Arnab, 2017; Lohr, 2021) have described methods for calculating variance for the HT and HH estimators. For the HT estimator, Barbiero and Mecatti (2010) proposed a variance calculation method using the bootstrap method to avoid complex calculations of inclusion probability. However, the use of the variance of the estimator would complicate the determination of central 95% CIs for load estimates because the distribution of load estimates would not be symmetrical, as in a normal distribution (Appling et al., 2015). Accordingly, we constructed CIs as interval estimates based on the following equation:

$$L_{HT} = \frac{1}{n}\sum_{i=1}^{n}\frac{l_i}{p_i} = \frac{1}{n}\sum_{i=1}^{n}\exp\left(\ln l_i - \ln p_i\right) \tag{12}$$

As a CI of L_{HT}, we calculate the bootstrap-t CI (Efron & Tibshirani, 1993) of the expectation of $\exp(\ln l_i - \ln p_i)$. Using the residual e_i ($= \ln l_i - \ln p_i$), a $100 \times (1 - \alpha)$ % central CI of L_{HT}, where α is the level of significance ($\alpha = 0.05$ in this study), can be calculated as follows. For convenience, x_i denotes $\exp(e_i) = \hat{l}_i/p_i$ (i = 1 to n) and the most right-hand side of Equation 12 is the mean of x_i. Let $\overline{x}$ be the mean of x_i. We construct the CIs of $\ln \overline{x}$ and then transform them back to $\overline{x}$ to avoid negative values of $\overline{x}$. We also assume that the distribution of $\ln \overline{x}$ is less skewed than that of $\overline{x}$. First, using n sample elements, the sample mean of $\ln \overline{x}$ and the jackknife estimate of the sample variance of $\ln \overline{x}$ are calculated as $\ln(\Sigma x_i/n)$ and $\boldsymbol{\Sigma}(\ln\overline{x}_{Ji} - \ln \overline{x})^2 \times (n - 1)/n$, respectively, where $\ln\overline{x}_{Ji}$ is the jackknife estimate of $\ln \overline{x}$ calculated from the $(n - 1)$ sample elements other than the ith sample. Next, B sets of a bootstrap replicated samples of x_i* with size n are generated by SRS with replacement from a sample of x_i of size n. For each bootstrap sample, the bootstrap sample mean $\ln\overline{x}^*$ and the jackknife estimate of the variance s^{*2} are calculated. The approximate pivot quantity t^* is defined as the pivot statistic $(\ln\overline{x}^* - \ln\overline{x})/s^*$ for each bootstrap sample, and B sets of t^* are sorted in ascending order. Finally, the upper and lower confidence limits for the two-sided $(1 - \alpha) \times 100\%$ CI are calculated as $\exp(\ln\overline{x} - s \times t^*_{B\alpha/2})$ and $\exp(\ln\overline{x} - s \times t^*_{B(1-\alpha/2)})$, respectively, where the ith order of t^* is denoted as t^*_i.

3. Data and Methods

3.1. Water Quality and Discharge Data

In this paper, we used WQ data sets from NCWQR (Heidelberg University, 2020) to test the validity of the methods for estimating annual loads for watersheds with various land uses and scales. Daily discharge data sets were downloaded from the USGS National Water Information System (NWIS, U.S. Geological Survey, 2016). Because the WQ data sets from the NCWQR sometimes contain several concentrations in a single day, we calculated daily mean concentrations following the procedure described in Appendix S1 by Hirsch (2014). We tested six WQ parameters with a sufficient number of observations: suspended solids (SS), total phosphorous (TP), soluble reactive phosphorous (SRP), nitrite, and nitrate (NO23), total Kjeldahl nitrogen (TKN), and chloride (Cl). We evaluated annual loads for years when the number of days with missing observations for any of the above six WQ parameters was less than 15 and the sum of the discharge on these missing days was less than 5% of the annual discharge. As a result, 25 combinations of years and watersheds were evaluated (Table 1). In total, 150 data sets (six WQ parameters × 25 combinations) were used for annual load estimation. These daily paired discharge-WQ data sets prepared by the authors can be downloaded from the Zenodo repository (https://doi.org/10.5281/zenodo.5154774).

Table 1
Sites and Years Used in the Load Evaluation

Station name	Abbrev.	In the watershed of	USGS gaging station number	Watershed area (km^2)	Land use (%)[a]					Target years (hydrological conditions)[a]
					A	P	F	U	O	
Blanchard	BLAN	Lake Erie	4,189,000	896	79	3	6	10	1	2008(d), 2009(*n*), 2016(*n*)
Cuyahoga	CUYA	Lake Erie	4,208,000	1,830	9	12	34	40	6	2007(d), 2016(w)
Great Miami	GREA	Ohio River	3,271,500	7,019	65	9	9	17	1	2007(d), 2008(*n*), 2014(*n*)
Honey	HONE	Lake Erie	4,197,100	386	81	2	10	7	1	1992(*n*), 2007(d), 2012(*n*), 2013(*n*)
Maumee	MAUM	Lake Erie	4,193,500	16,388	73	6	6	11	3	2013(w)
Muskingum	MUSK	Ohio River	3,150,000	19,215	24	19	43	12	2	2007(*n*), 2008(vw), 2011(w), 2012(d), 2013(vd), 2015(d)
Portage	PORT	Lake Erie	4,195,500	1,108	84	1	4	9	1	2019(w)
Rock	ROCK	Lake Erie	4,197,170	90	72	8	11	9	0	2008(w), 2012(*n*), 2015(d)
Tiffin	TIFF	Lake Erie	4,185,000	1,061	61	15	9	7	8	2017(*n*), 2019(vw)

Note. [a]The information on the U.S. sites is from the Heidelberg University National Center for Water Quality Research (https://ncwqr.files.wordpress.com/2020/01/basic-station-information-ao-wy2019.xlsx). A; Agriculture, P; Pasture, F; Forest, U; Urban, and O; Others.

[a]Hydrological conditions; (vd) very dry (0th to 10th percentile), (d) dry (10th to 33rd percentile), (n) normal (33rd to 66th percentile), (w) wet (66th to 90th percentile), and (vw) very wet (90th to 100th percentile) among 30 to 51 water years (1971 to 2021 for Blanchard, Cuyahoga, Maumee, Portage, and Tiffin, 1992 to 2021 for Great Miami, 1977 to 2020 for Honey, 1971 to 2021 for Muskinghum, and 1984 to 2021 for Rock).

The hydrological conditions of the data sets are also listed in Table 1. For each watershed, years with complete USGS daily discharge were selected from 1971 to 2021. The number of years with complete discharge data ranged from 30 to 51, depending on the watershed (see Table 1). Using annual discharge values calculated from in years with complete discharge data for each watershed, the hydrological condition in each year was classified as follows. The years with 0 to 10th, 10th to 33rd, 33rd to 66th, 66th to 90th, and 90th to 100th percentiles of annual discharge were classified as very dry (vd), dry (d), normal (n), wet (w), and very wet (vw) years, respectively. In the above classification, the lower limit was not included. As a result, among the 25 combinations of years and watersheds, one was very dry, seven were dry, 10 were normal, five were wet, and two were very wet. Overall, the evaluated data sets covered wide hydrological conditions.

3.2. The Evaluation Method for the HT Estimator

3.2.1. Evaluation of Load Estimates Based on Samples Obtained With RDM, RHF, and FP Sampling

Using 150 data sets, we evaluated the bias of load estimates and the coverages of CIs, which are the probabilities of CIs bracketing the true load value. The number of sample elements collected by RDM or FP sampling (RDM or FP samples) was set to 12 per year. The number of elements collected by RHF sampling (RHF samples) was 20, that is, 12 by RDM sampling and eight by high-flow sampling based on the high-flow threshold of q_{90}. The details of the sampling procedures were described in Section 2.2, and none of the sampling procedures involved duplicate elements. With the Monte Carlo method, M sets of a sample of size n were withdrawn from the population for each sampling strategy ($M = 20{,}000$ in this study). During each sampling stage for elements from the population, sampling without replacement was performed as described in Section 2.2. Based on M sets of point estimates and the CIs, pBIAS in Equation 13 and the coverage were calculated and used to evaluate the performance of the proposed methods.

$$\text{pBIAS} = 100 \times \left(\frac{1}{M} \sum_{i=1}^{M} L_{HT} - L_{TRUE} \right) \Big/ L_{TRUE} \tag{13}$$

The unbiasedness of load estimates was verified based on whether pBIAS values were almost zero, and the appropriateness of the interval estimation approach was evaluated according to whether the coverages of the 95% CIs were approximately 0.95. The number of bootstrap replicates B used in configuring CIs was set to 2,000 in this study. Although pBIAS values produced by the HT estimator are expected to be zero, the coverage of 95%

bootstrap-t CIs is not always 0.95 because coverages are influenced by the distribution of the population; for example, the coverages of bootstrap-t CIs decrease from the assumed confidence level for a very skewed population distribution when the sample size is small (Owen, 1992). We also employed decision tree analysis based on the classification and regression tree (CART) algorithm to investigate the factors affecting the performance of 95% CI coverages.

In evaluating the validity of the proposed methods, we used the true load value L_{TRUE}, which is the sum of observed daily loading rates during a year. When a data set included missing observations, we defined the population data by excluding those missing values from the data set. The value of the true load calculated from daily paired discharge and WQ data could be characterized by substantial bias compared to the true load value defined based on the total sum of loading rates with a much shorter time step (Lee et al., 2017). However, for the purpose of evaluating the proposed methods, we can use the daily time step data. Accordingly, when the proposed methods are applied to estimate the true load defined based on a short time step, such as an hourly time step, discharge data must be collected in this short time step, and WQ observations must be paired to discharge values accordingly, even for rivers draining large watersheds.

In load estimation and resampling simulation, proprietary FORTRAN codes were used. Using the Mersenne Twister algorithm with a period long enough for the Monte Carlo evaluation (Matsumoto & Nishimura, 1998), pseudorandom numbers were generated.

3.2.2. Evaluation of Uncertainty Reduction by IPPS RS

In uncertainty reduction with IPPS RS, as described in Section 2.3, the minimum size of a resampled sample was set to four because bootstrap-t CIs require at least four sample elements to avoid confidence limits of infinity (Owen, 1992). To minimize the chance of failed resampling with fewer than four sample elements, IPPS RS was implemented M_{RS} times repeatedly for each sample, and the resampled sample with the maximum size was adopted. M_{RS} was set to 1,000 in this study. When the maximum size of resampled samples was still less than four after M_{RS} trials, IPPS RS was regarded as failing. However, the HT estimator based on a resampled sample through M_{RS} retrials may not provide unbiased load estimates because the sampling probability of a resampled sample is no longer the same as the expected probability p_1 in Equation 10. A change in sampling probability from the expected p_1 could result from a small-size sample with different statistical properties than those of the population.

Consequently, the actual sampling probability after M_{RS} resampling trials for the population element, p_j ($j = 1$ to N), must be calculated retrospectively with the Monte Carlo method. In the first step of the p_j calculation, M_M sets of a sample of size n were withdrawn from the population based on a sampling strategy with probability p_0 (in this study, $M_M = 200{,}000$). In each sampling step, sample elements were collected without replacement. Next, for each sample, IPPS RS with sampling probability p_1 based on Equation 10 with the specified β_1 value was conducted. Then, when the sample size was larger than three after M_{RS} trials, the number of resampling instances, d_j, was determined for M_M sets of a sample. The posterior sampling probability p_j was then numerically calculated as:

$$p_j = d_j \bigg/ \sum_{i=1}^{N} d_i \tag{14}$$

This sampling probability p_j can provide unbiased load estimates based on Equation 3. We also used this posterior sampling probability in load estimation based on FP samples. It should be noted that a narrative sampling procedure must be translated into an algorithm that can be coded to calculate sampling probabilities with the Monte Carlo method.

Using the IPPS RS described above, the effectiveness of the uncertainty reduction by IPPS RS for RDM, RHF, and FP sampling strategies was evaluated. Load estimation based on IPPS RS was performed for β_1 values ranging from 0.1 to 3.0 in increments of 0.1 because sampling probability p_{1_i} based on a simple rating curve is determined only by the β_1 value (Equation 10). As the zero β_1 value in Equation 9 corresponds to complete dilution by discharge and random sampling in time, an evaluation of IPPS RS started with β_1 of 0.1. Additionally, because β_1 values greater than 3.0 would reduce the resampled sample size and increase the uncertainty of load estimates, we set the maximum β_1 value for the evaluation at 3.0. We consider that this maximum β_1 value is appropriate even for particulate WQ parameters having high β_1 values such as SS and TP when samples by random sampling in time or flow-proportional sampling have been collected. With this estimation procedure, n_e sets, the number of successful IPPS RS results among 30 sets of β_1 values for RS, of load estimates were calculated from a sample. As a result, (n_e+1) load estimates, including the load estimates based on the original sample with sampling probability p_0, were calculated for a sample. However, the load estimates with the minimum-width CIs should not be selected from (n_e+1) estimates because such an estimate tends to provide much lower coverage than the specified confidence level. This is because each CI has a 5% probability of not bracketing $L_{TRUE,}$ and a narrow CI tends to have a comparatively larger lower confidence limit and smaller upper confidence limit. To choose a narrow CI while avoiding a reduction in coverage, the narrowest CI within the range in which lower confidence limits are not larger than L_{TRUE} was selected as a candidate for the most precise interval estimate. That is, the $n_e \times (1 - \alpha/2)^{n_e}$ th CI width in descending order of the n_e CI widths based on IPPS RS was compared to the CI width derived from an original sample collected with sampling probability p_0. The load estimate with the narrower CI of the above two was selected as the most precise, that is, the reduced-uncertainty load estimate.

4. Results

4.1. Process of Uncertainty Reduction With IPPS RS

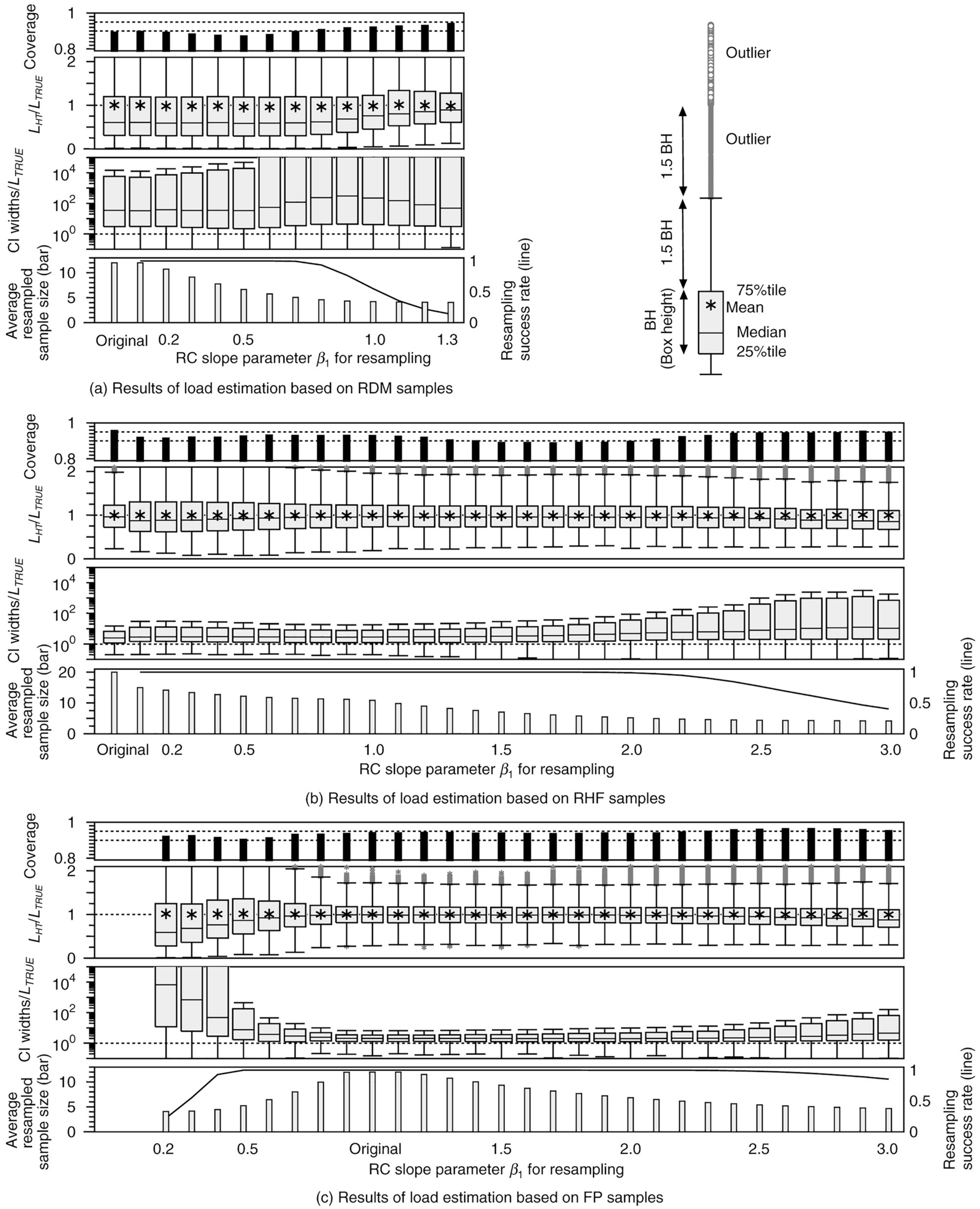
Figures 2a–2c give the results of load estimation for the BLAN2008 SS data set obtained through IPPS RS based on RDM, RHF, and FP samples, respectively. The results include the distributions of point estimates and CI widths, the coverages of CIs, the original sample size and average resampled sample sizes, and the resampling success rates; this variable is defined as the ratio of the number of successfully obtained IPPS RS samples to the number of Monte Carlo evaluations M. The point estimates and CI widths in Figure 2 are standardized based on L_{TRUE}. Regression with Equation 9 based on the population of this data set yields a β_1 value of 1.523. Figure 2 shows that load estimates are unbiased for all β_1 values used in IPPS RS with any sampling strategy tested here. The CI widths do not always decrease as β_1 approaches 1.5. FP sampling provides narrower CIs at $\beta_1 = 1.0$, the value of sampling, than at $\beta_1 = 1.5$, the value of the population. This difference is associated with the sample size reduction due to resampling. In RDM sampling, IPPS RS gives the narrowest CI width based on the original sample elements. Additionally, the CI width derived from an original sample was narrower in RHF and FP sampling. These results imply the dominant effect of sample size reduction on load estimation. IPPS RS does not seem to reduce the uncertainty of load estimates on average, although the uncertainty reduction can occur in load estimation for each sample.

4.2. Effectiveness of the HT Estimator in Load Estimation

The results of load estimation for the SS and NO23 data sets based on RDM samples are shown in Figures 3a and 3b, respectively. The results include the coverages of CIs and the distributions of both point estimates and CI widths. Point estimates and CI widths are standardized based on L_{TRUE}, and the hydrological conditions of data sets are also shown. Figures 4 and 5 give the results of these two WQ parameters based on RHF sampling and FP sampling, respectively. The results of load estimation for the other WQ parameters are shown in Figures S2 to S7 in Supporting Information S1. These figures also give the uncertainty reduction results for IPPS RS. The values of pBIAS and coverages for all the data sets are listed in Tables S1 to S3 in Supporting Information S1.

In these figures, the HT estimator L_{HT} can provide unbiased estimates for any WQ parameter, any watershed, any water year, and any sampling strategy tested here. Additionally, in Tables S1 to S3, the maximum values of absolute pBIAS among 150 data sets for RDM, RHF, and FP sampling are 1.4%, 0.3%, and 1.5%, respectively. These results confirm that L_{HT} can provide unbiased load estimates as long as the sampling probability is properly defined. The relationship between the estimation results and hydrological conditions of the data sets is not clear because the estimates were standardized by L_{TRUE}. In summary, a proper sampling probability must be determined theoretically based on the procedure in Section 2.2 or retrospectively by the Monte Carlo method, as discussed in Section 3.2.2, to calculate unbiased L_{HT} values for existing monitoring data collected with an operating sampling strategy.

These figures also show that the CI widths of estimates based on RDM samples (Figure 3) were so wide that meaningful interpretation of the estimates could not be made, especially for particulate matter with this sample size. The CI widths based on FP samples (Figure 5) were narrowest among those for the three sampling methods.

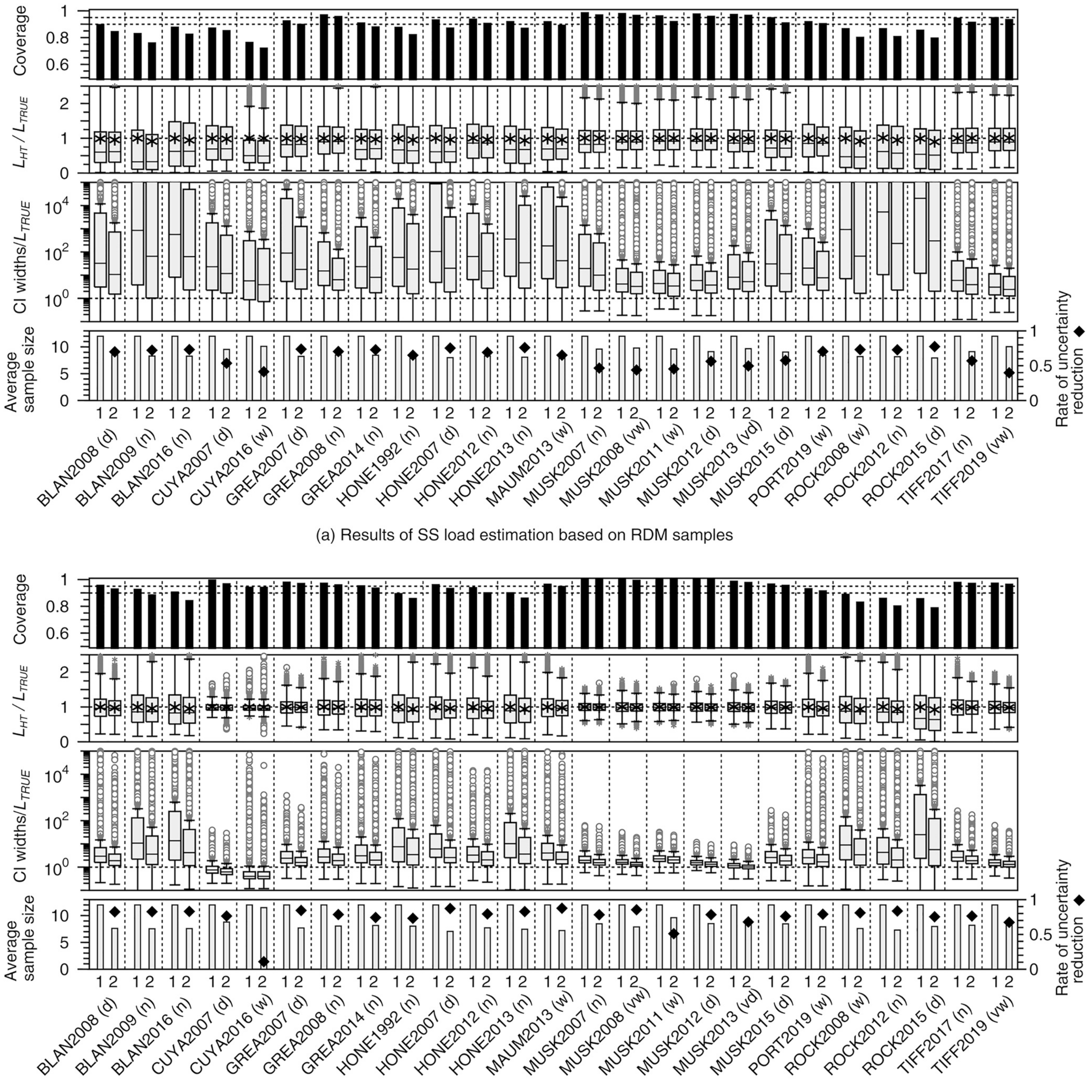


(a) Results of load estimation based on RDM samples

(b) Results of load estimation based on RHF samples

(c) Results of load estimation based on FP samples

Figure 2. The results of load estimation for the BLAN2008 suspended solids data set with inclusion probability proportional to size random sampling.

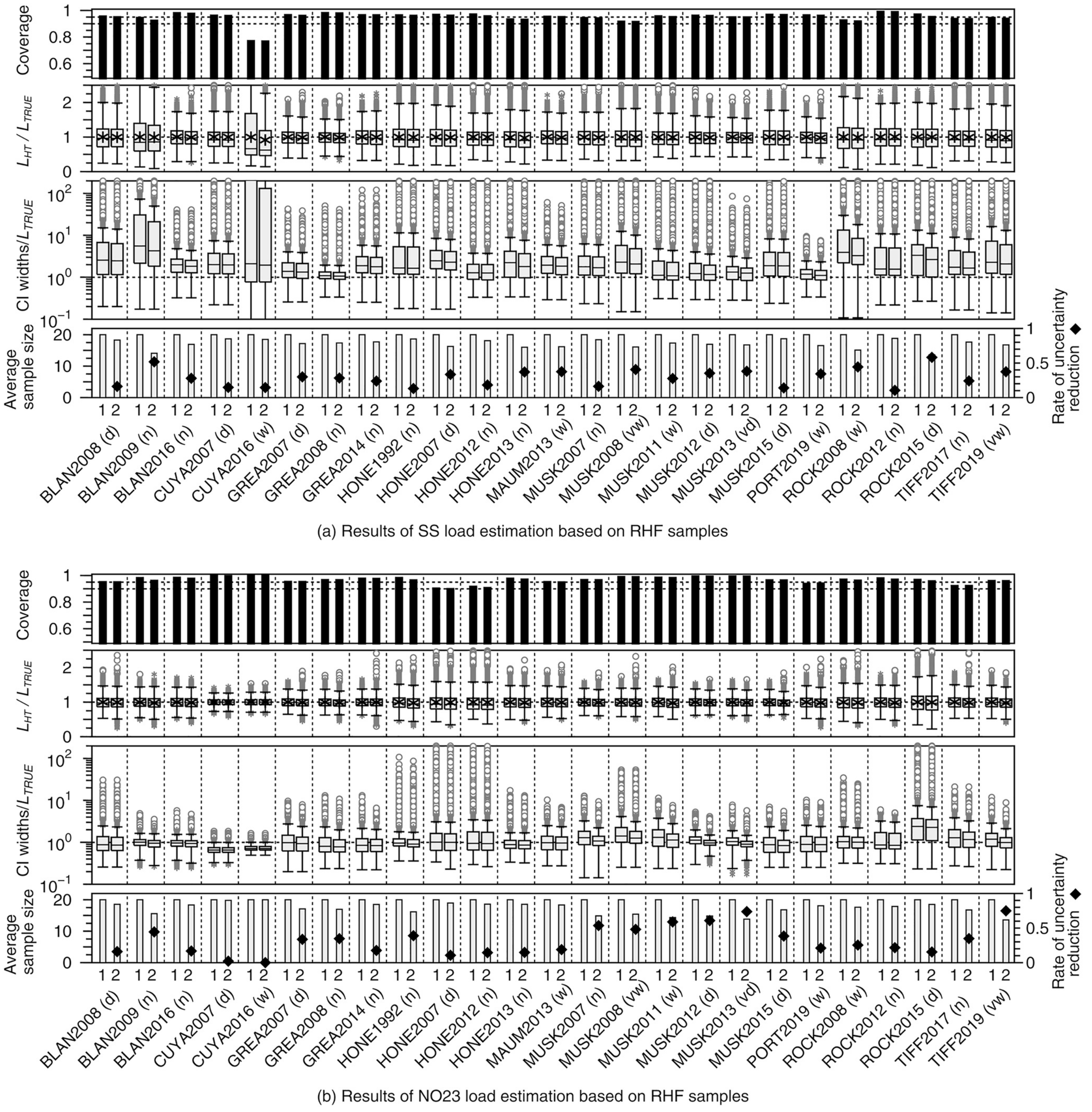


(a) Results of SS load estimation based on RDM samples

(b) Results of NO23 load estimation based on RDM samples

Figure 3. Results of load estimation for the suspended solids and NO23 data sets based on monthly random sampling samples. The numbers one and two on the horizontal axis correspond to the estimation results based on the original sample elements and with uncertainty reduction, respectively. The abbreviations (vd), (d), (n), (w), and (vw) represent the hydrological conditions of very dry, dry, normal, wet, and very wet, respectively.

The CI widths based on RHF samples (Figure 4) were much narrower than those based on RDM samples, indicating the large effect of high-flow samples on the reduction in uncertainty. Low coverages less than 0.85 were sometimes found in the estimation results based on RDM samples (18 cases out of 150) but rarely in the results based on RHF and FP samples (five and four cases, respectively). Most of the results exhibited coverages greater than or equal to 0.90 (108, 137, and 142 cases out of 150 for RDM, RHF, and FP sampling, respectively). Overall, the proposed interval estimation method provides appropriate CIs for load estimates, except for estimates based on RDM samples.

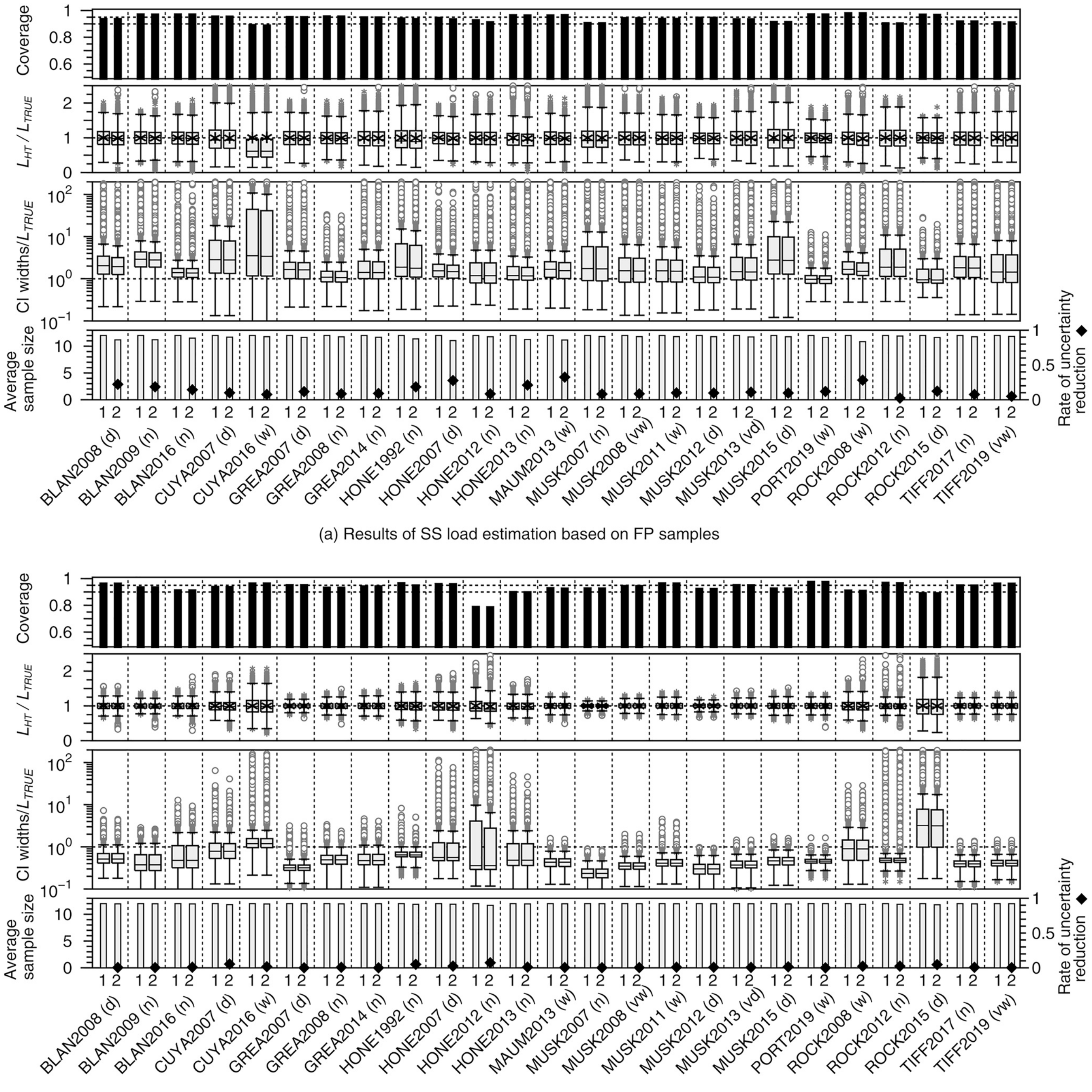


(a) Results of SS load estimation based on RHF samples

(b) Results of NO23 load estimation based on RHF samples

Figure 4. Results of load estimation for the suspended solids and NO23 data sets based on RHF samples. The numbers one and two on the horizontal axis correspond to the estimation results based on original sample elements and with uncertainty reduction, respectively. The abbreviations (vd), (d), (n), (w), and (vw) represent the hydrological conditions of very dry, dry, normal, wet, and very wet, respectively.

To investigate the major factors leading to low coverages less than 0.85 in the estimation results in RDM sampling, decision tree analyses based on the CART algorithm were performed using R (version 4.1.0) and its packages rpart, rpart.plot, and partykit. It was expected that some of the properties of the natural logarithmic distribution of the observed loading rates, lnl_j ($j = 1$ to N), caused these low coverages. In the decision tree analysis, the target variable of the tree was the status value of the coverage, that is, 0, 1, 2, or 3 for coverages ranging from 0.9 to 1.0, from 0.85 to 0.9, from 0.80 to 0.85, or less than or equal to 0.80 (lower limit not included), respectively. The input variables were l_{max}/L_{TRUE}, l_{max}/l_{max2}, the maximum normalized variable of lnl_j (Z_{max}), the variance of lnl_i (s^2), the



(a) Results of SS load estimation based on FP samples

(b) Results of NO23 load estimation based on FP samples

Figure 5. Results of load estimation for the suspended solids and NO23 data sets based on flow proportional samples. The numbers one and two on the horizontal axis correspond to the estimation results based on original sample elements and with uncertainty reduction, respectively. The abbreviations (vd), (d), (n), (w), and (vw) represent the hydrological conditions of very dry, dry, normal, wet, and very wet, respectively.

skewness of $\ln l_j$, and the kurtosis of $\ln l_j$, where l_{max} is the maximum value of l_j, l_{max2} is the second largest l_j value, the jth normalize variable $Z_j = (\ln l_j - \overline{\ln l_j})/s$, $s^2 = \Sigma(\ln l_j - \overline{\ln l_j})^2/(N-1)$, and $\overline{\ln l_j} = \Sigma \ln l_j/N$. The results of the decision tree analysis (Figure S8 in Supporting Information S1) indicate that coverage tends to be less than 0.85 when $Z_{max} > 3.42$, and the other input variables have little effect on low coverage; extremely large daily loading rates caused the low coverages of estimates based on RDM samples. However, the existence of such extreme values cannot be detected in small sample sets, such as sets with 12–20 elements. In this regard, another similar decision tree analysis was performed while the target variable was held constant. The input variables for this

Table 2
Average Values, Standard Deviations (SD), Minimum Values, and Maximum Values of pBIAS (%) and the Coverage of Load Estimates for 150 Data Sets

	RDM			RHF			FP		
	Org	Red	Red-Org	Org	Red	Red-Org	Org	Red	Red-Org
					pBIAS				
Average	0.0	−3.1	−3.1	0.0	−1.6	−1.6	0.0	−0.6	−0.6
SD	0.4	2.5	2.1	0.2	1.0	0.8	0.2	0.8	0.6
Minimum	−1.1	−10.1	−9.0	−0.8	−8.3	−7.5	−0.6	−3.5	−2.9
Maximum	1.4	0.8	−0.6	0.4	0.7	0.3	0.6	1.3	0.7
					Coverage				
Average	0.93	0.90	−0.03	0.96	0.96	0.00	0.95	0.94	0.01
SD	0.05	0.07	0.02	0.03	0.03	0.00	0.03	0.03	0.00
Minimum	0.76	0.72	−0.04	0.77	0.77	0.00	0.78	0.78	0.00
Maximum	1.00	1.00	0.00	1.00	1.00	0.00	0.99	0.99	0.00

Note. Org, Red, and Red-Org represent load estimation results based on the original sample elements, the results after uncertainty reduction, and the difference between Red and Org.

analysis were q_{max}/Q_T, the maximum normalized variable of $\ln q_j$ (Z_{maxq}), the variance of $\ln q_j$ ($s_q{}^2$), the skewness of $\ln q_j$, and the kurtosis of $\ln q_j$, where the normalized variable $Z_{qj} = (\ln q_j - \overline{\ln q_j})/s_q$, $s_q{}^2 = \Sigma(\ln q_j - \overline{\ln q_j})^2/(N - 1)$, and $\overline{\ln q_j} = \Sigma \ln q_j/N$. The result of this analysis (Figure S9 in Supporting Information S1) suggests that coverage tends to be less than 0.90 when $Z_{maxq} > 2.82$. In brief, the CIs of load estimates based on RDM samples tend to have low coverage when an extremely large value of daily discharge exists.

In contrast to the load estimation results in RDM sampling, the coverages of estimates based on RHF and FP samples have appropriate values of approximately 0.95 (0.90 to 1.00), even for data sets with extremely large values of l_i, because the variances of e_i in Equation 12 for RHF and FP samples are much smaller than those for RDM samples. The results of RHF sampling suggest that the current WQ monitoring strategy employed in the CBP could provide unbiased estimates and proper CIs with the HT estimator if the difference between estimates based on samples obtained by random and systematic sampling is negligible.

4.3. Effectiveness of Uncertainty Reduction

Figures 3–5 and Figures S2 to S7 in Supporting Information S1 also provide the results of load estimation after uncertainty reduction. The rate of uncertainty reduction is shown only for IPPS RS in these figures. This rate is defined as the ratio of the number of cases in which reduced CI widths are adopted and the resampled sample size is smaller than the original sample size to the total number of evaluations M. The rate of uncertainty reduction is high in RDM sampling and low in FP sampling. These results show that the reduction in uncertainty is significant only for RDM samples, and the effect of reduction is limited for RHF and negligible for FP samples. In addition, even for RDM samples, the magnitude of the reduction in CI width is not satisfactory, and the reduced CI width remains wide. Because both RHF and FP sampling involve the collection of high-flow samples, the effect of IPPS RS on uncertainty reduction is small for these two sampling strategies. In these strategies, the adverse effect of size reduction in a resampled sample tends to outweigh the positive effect of sampling probability modification. Table 2 shows the average values, standard deviations, minimum values, and maximum values of pBIAS and the coverages out of 150 data set evaluations; additionally, the difference between the values based on the original sample and the sample after uncertainty reduction is shown for the three sampling strategies. From this table, IPPS RS leads to a slight increase in both absolute pBIAS and its standard deviation. IPPS RS seems to have no adverse effect on coverages. We consider this limitation of applying IPPS RS to be acceptable.

In summary, IPPS RS can provide narrower CIs than other methods with only slightly biased estimates, and the adoption of uncertainty reduction is effective, mainly in load estimation based on RDM samples. Although IPPS

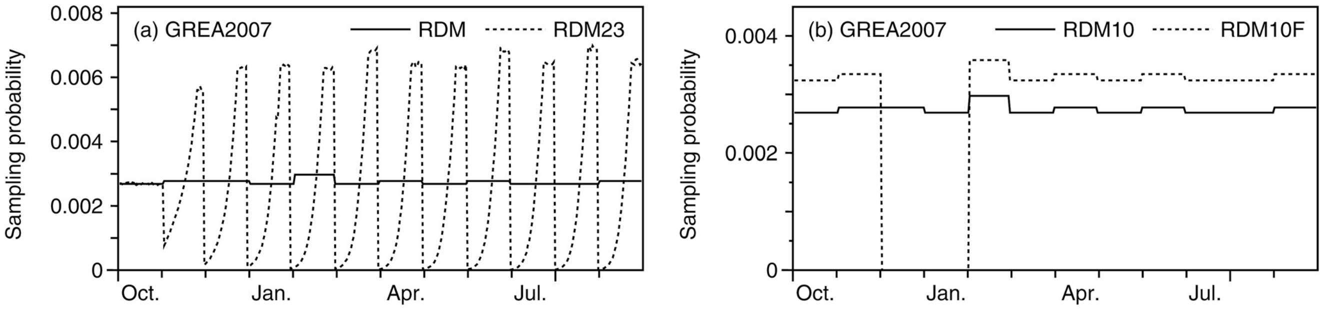


Figure 6. Sampling probabilities for monthly random sampling (RDM), RDM23, RDM10, and RDM10 F sampling strategies for the GREA2007 suspended solids data set. Figure 6a shows the probabilities for RDM and monthly sampling with a minimum interval of 23 days (RDM23). Figure 6b provides the probabilities for the sampling strategies that remove two sample elements randomly from monthly random samples (RDM10) and remove two sample elements in December and January from monthly random samples (RDM10 F).

RS can reduce the CI width based on RDM samples, the magnitude of reduction in the CI width when IPPS RS is applied is much smaller than that obtained by changing the sampling strategy from RDM to RHF or FP sampling.

5. Discussion

5.1. Effects of the Sampling Probability on Load Estimates

The question of whether the proposed load estimation methods are applicable to existing WQ samples is posed because most sampling strategies in operation are neither systematic nor random, as evaluated above. For example, a high-flow sampling strategy may not always collect the predefined number of sample elements in a very dry year or may finish collecting the predefined number of elements much earlier than expected in a very wet year. In such cases, the true randomness of high-flow sample elements may be uncertain. To determine whether the actual sampling probability p of sample elements follows a certain distribution, only the nonconformity of the actual p values to a certain distribution can be statistically tested by a goodness-of-fit test. For example, Anderson–Darling statistics can be used to test the null hypothesis that cumulative sampling probability scores of Σp for sample elements based on a certain probability distribution follow a uniform distribution at a certain significance level (D'Agostino & Stephens, 1986). However, it is impossible to retrospectively test the conformity of actual p values for an existing WQ sample to a certain distribution because we can only choose to accept the alternative hypothesis of nonconformity of actual p values to a certain distribution or to decline that alternative hypothesis based on a statistical test. In practice, the best way to determine the actual sampling probability of a sample would be to simulate sampling probabilities with the Monte Carlo method described in Section 3.2.2 based on the employed sampling strategy at WQ monitoring sites.

In load estimation with the HT estimator, the definition of the actual sampling probability p of sample elements is most important. We explain this importance with an example of calendar-based sampling or random sampling. Here, we consider the difference between load estimates based on RDM samples and quasi-systematic sampling with a minimum interval, as tested by Lee et al. (2019). In quasi-systematic sampling, one sample element is collected randomly each month through RDM sampling, but the interval between element collection dates must be at least 23 days (hereafter, this sampling procedure is referred to as RDM23 sampling). Figure 6a provides the simulated sampling probability for RDM23 sampling with the Monte Carlo method for the GREA2007 SS data set. In RDM23 sampling, the sampling probability in the former half of a month gradually approaches zero (but does not reach zero) based on the minimum interval of 23 days.

To investigate the effect of the choice of sampling probability on load estimates, we evaluated load estimates calculated based on the following three combinations of the sampling probabilities used by the HT estimator and employed in the collection of sample elements: RDM and RDM (R-R), RDM and RDM23 (R-R23), and RDM23 and RDM23 (R23-R23). Table 3 provides the average values of pBIAS and coverages of the estimation results for 25 combinations of watersheds and years for each WQ parameter and the above three combinations. Individual values of pBIAS and coverage are also listed in Tables S4a to S4f in Supporting Information S1. These results show that the coverages of R-R23 are appropriate and similar to those of R-R, but the pBIAS values of R-R23 are not zero. For R23-R23, the pBIAS values are almost zero, but the coverages decrease from the confidence

Table 3
Average Values of pBIAS (%) and Coverage (CVR) for Combinations of RDM Sampling and RDM23 Sampling Probabilities

	RDM-RDM		RDM-RDM23		RDM23-RDM23	
WQ parameters	pBIAS	CVR	pBIAS	CVR	pBIAS	CVR
SS	0.0	0.89	−14.5	0.87	−1.2	0.79
TP	0.2	0.93	0.3	0.94	−0.4	0.79
SRP	0.0	0.92	−7.7	0.94	0.0	0.87
NO23	0.1	0.99	−4.4	1.00	0.1	0.75
TKN	0.0	0.94	−8.3	0.94	−0.3	0.79
Cl	0.1	0.91	−15.7	0.90	−0.7	0.79

level of 0.95. Because both R-R and R23-R23 used the same sampling probability for the collection of sample elements and for HT estimator calculations, the resulting pBIAS values are almost zero. The decreased coverages for R23-R23 could be attributed to the near-zero sampling probability in the former half of each month. Nevertheless, should we avoid the combination of R-R23 in load estimation because different sampling probabilities are considered in element collection and load calculations?

In practice, the combination of R-R23 provides a particular subset of load estimates based on the combination of R-R (see Figure S10 in Supporting Information S1) because the RDM23 sampling approach involves the collection of a subset of RDM samples. As a result, load estimates based on the samples obtained with RDM23 sampling have appropriate coverages when RDM sampling probabilities are used to calculate L_{HT}, although this combination of R-R23 would not provide zero pBIAS. We should note that the bias observed for R-R23 is not due to the improper combination of sampling probabilities but results from the limitation of an application of the Monte Carlo method to a population with a small size, such as 365. Considering the low coverages provided by R23-R23, we should choose R-R23; notably, a proper evaluation of uncertainty is more important than an individual point estimate when only one sample can be used for load estimation. In summary, we should employ a random sampling probability in calculating L_{HT}, even for samples collected with a constrained random sampling method, such as RDM23 or systematic sampling. We should not use a sampling probability that yields near-zero probability values for some elements of the population, even though the goal of sampling is to collect sample elements with almost equal sampling probabilities.

5.2. Implications for Missing Observations and Type-I Censored Data

WQ monitoring is often associated with missing observations. For example, one sample element in RDM sampling may not be collected due to inaccessibility to the monitoring site for some reason. In such cases, we cannot use the sampling probability in RDM sampling with a sample size of 12 to calculate L_{HT}. Additionally, the sampling probabilities with missing observations in fixed-month sampling provide biased load estimates due to the zero probability of sampling for some elements of the population (nonsampling error). To solve the problem caused by missing observations, we consider two sampling strategies: the strategy of removing two sample elements randomly from RDM samples (RDM10) and the strategy of removing two elements in the fixed months of December and January from RDM samples (RDM10 F). These 2 months were selected because they have approximately the highest and the second highest average daily discharge in a year (Figure S11) and removing the elements from these 2 months would have a great impact on the load estimation bias. The sampling probabilities for these two sampling strategies are shown in Figure 6b. In RDM10 F sampling, the sampling probability becomes zero in December and January. In these two sampling scenarios with missing observations, we assumed that there were no missing discharge observations.

Using these two strategies, we evaluated load estimates based on the following three combinations of sampling probabilities used in load calculation and employed for the collection of sample elements: RDM10 and RDM10 (R10-R10), RDM10 and RDM10 F (R10-R10 F), and RDM10 F and RDM10 F (R10F-R10 F). Table 4 provides the average values of pBIAS and the coverages of the estimation results for 25 combinations of watersheds

Table 4
Average Values of pBIAS (%) and Coverage (CVR) Based on Combinations of RDM10 and RDM10 F Sampling Probability

WQ parameters	RDM10-RDM10		RDM10-RDM10F		RDM10F-RDM10F	
	pBIAS	CVR	pBIAS	CVR	pBIAS	CVR
SS	−0.1	0.90	−4.9	0.90	−20.8	0.88
TP	0.1	0.91	−8.8	0.90	−24.1	0.88
SRP	−0.2	0.91	−9.8	0.90	−24.8	0.86
NO23	0.2	0.94	−9.0	0.93	−24.1	0.87
TKN	0.0	0.92	−6.5	0.91	−22.1	0.89
Cl	0.1	0.95	−7.9	0.94	−23.2	0.89

and years for each WQ parameter and the three combinations of sampling probabilities. Individual values of pBIAS and coverage are also listed in Tables S5a to S5f in Supporting Information S1. These results indicate that R10-R10 provides unbiased estimates and appropriate coverages for CIs, as shown in Table 3, but R10F-R10 F yields biased estimates and poorer coverages than R10-R10. R10-R10 F provides biased estimates but appropriate coverages. Using nonsampling probabilities when calculating L_{HT} results in biased estimates and poor coverage of CIs. Consequently, to avoid the nonsampling error, we should not fix the times of the missing observations. In addition to that for the cases of constrained random sampling, the apparent bias for R10-R10 F is not from an improper combination of sampling probabilities but results from the limitation of applying the Monte Carlo method to a population with a small size.

The above results demonstrate the adverse effect of nonsampling error and the importance of avoiding a zero sampling probability for elements in a population when calculating L_{HT}. Such nonsampling error could also be caused by high-flow sampling if samples were only collected for daily discharges above a certain threshold. In this case, load estimates would be characterized by bias and low CI coverage because low flow-related elements would have zero sampling probability. To calculate unbiased load estimates and appropriate CIs, low-frequency calendar-based sampling should be employed together with high-flow sampling to avoid nonsampling error, as in RHF sampling. In addition, this technique of avoiding nonsampling error for proper uncertainty estimation is applicable to snowpack dominated systems that may cause missing observations due to inaccessibility to monitoring sites during winter, although pBIAS would slightly increase.

If a data set includes observations below the relevant detection limit, that is, Type-I censored data, the calculation of L_{HT} with Equation 3 will be impossible because both l_i and c_i will be unknown for elements under the detection limit. In this case, the Tobit model (Tobin, 1958) can be applied to derive an RC for Type-I censored data, and the resampled residuals of detected elements can be added to the load estimates with the RCM to estimate l_i or c_i for the sample elements below the detection limit. Using these extrapolated values, L_{HT} can be calculated.

5.3. Re-Evaluation of WQ Monitoring Strategies and Improvements to Load Estimates Based on RCM Using IS

Here, we describe a method for re-evaluating operating sampling strategies employed at WQ monitoring sites based on the HT estimator and historical data. A WQ monitoring program aimed at pollutant load estimation may assume a certain level of uncertainty in estimates according to the purpose of the project, for example, load estimates with a 20% coefficient of variance or a certain value of CI width. In this case, the best sampling procedure and sample size must be determined to satisfy the specified uncertainty level for load estimates in the project. The appropriate sample size considering uncertainty could be derived by simulating load estimates using historical discharge data if the sampling procedure is determined and both RC parameters and variance of RC residuals s_l^2 based on historical data are available. In this situation, we can synthesize the loading rates l_{sj} ($j = 1$ to N) as follows:

$$l_{sj} = \hat{l}_j \times \exp(s_l \times nrnd_j) \tag{15}$$

where $\hat{l}_j$ is the estimated loading rate for the jth sample calculated with an RC using the discharge value and $nrnd_j$ is a normal random number following $N(0, 1)$. Using l_{sj} and q_j, we can calculate the distributions of CIs for L_{HT} with the Monte Carlo method based on various sample sizes under the assumed sampling procedure. Then, we can choose the appropriate sample size based on the median CI value or the CI value at a certain percentile. If there is no feasible sample size that can support the appropriate level of uncertainty, the sampling procedure must be redesigned or the target level of uncertainty reset.

The proposed methods also improve the load estimates obtained with the RCM using IS (Tada & Tanakamaru, 2021) because the simulated sampling probabilities used in the Monte Carlo method described in Section 3.2.2 can eliminate the maximum sample size limitation. In the RCM using IS, the maximum sample size for PPS samples was introduced to avoid duplicate sample elements, which lowers the coverage of the CIs of estimates because PPS sampling is a form of UPSWR sampling. This limitation prevented a reduction in the uncertainty of load estimates by increasing the sample size. However, the posterior calculation of the sampling probability enables IPPS sampling (UPSWOR sampling) at a much larger sample size than the maximum size in the RCM using IS. In this case, as approximate IPPS sample elements, only independent sample elements are chosen from the sample elements obtained through PPS sampling with duplicate elements. This procedure would provide more precise load estimates than those based on PPS samples with the maximum size limitation because an increase in sample size would effectively reduce the uncertainty of load estimates.

6. Conclusions

Based on small-size samples collected with simplified sampling strategies similar to those commonly used at WQ monitoring sites, we applied the Horvitz-Thompson estimator for annual load estimation and verified the unbiasedness of the estimator for various data sets. We also emphasized the importance of properly defining the sampling probability and demonstrated a posterior method of calculating the sampling probability with the Monte Carlo method. De Vries and Klavers (1994) emphasize the “monitoring strategy first, the calculation methods second” because the magnitude of load estimates is largely determined not by the calculation method used but by the monitoring strategy. In regard to unbiased load estimation with respect to the sampling strategy and an LCM, the saying should be “monitoring strategy first, calculation method fixed”. Based on the HT estimator, we can also construct appropriate CIs and derive unbiased load estimates. It is important to choose an appropriate sampling probability for unbiased load estimation and proper uncertainty estimation. For example, sampling probability based on random sampling can provide the proper uncertainty estimation for systematic samples. We can also calculate accurate load estimates from a sample with missing observations by avoiding zero sampling probability based on reduced sample size. Since unbiased load estimation can be performed universally with the HT estimator, as demonstrated in this study, the use of a biased estimator would require specific justification. Based on the proposed methods, water pollution control measures could be refined, the effects of pollutant load reduction could be assessed, WQ monitoring strategies could be redesigned, and unbiased load estimates could be obtained from existing WQ monitoring data.

List of Abbreviations

BCF	bias correction factor
CBP	Chesapeake Bay program
CI	confidence interval
FP sampling	flow-proportional sampling
HH estimator	Hansen-Hurwitz estimator
HT estimator	Horvitz-Thomson estimator
IPPS	inclusion probability proportional to size
IPPS RS	rejection sampling based on IPPS
LCM	load calculation method
MVUE	minimum variance unbiased estimator
pBIAS	percent bias
PPS	probability proportional to size
QMLE	quasimaximum likelihood estimator
RC	rating curve
RCM	rating curve method
RDM sampling	monthly random sampling
RHF sampling	monthly random sampling plus high-flow random sampling
RS	rejection sampling
SALT	selection at list time
SIR	sampling/importance resampling
SRS	simple random sampling

UPSWOR	unequal probability sampling without replacement
UPSWR	unequal probability sampling with replacement
WQ	water quality

Data Availability Statement

The data used in this paper are currently available from Heidelberg University's National Center for Water Quality Research site (https://ncwqr-data.org/HTLP/Portal) and the USGS National Water Information System (http://waterdata.usgs.gov/nwis/ or https://doi.org/10.5066/F7P55KJN). The actual daily paired discharge and WQ data used in this paper are available from the Zenodo repository (https://doi.org/10.5281/zenodo.5154774).

Acknowledgments
This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Nos. JP23580333, JP15K07646, and JP21K05828. The authors acknowledge the provision of the hydrological and water quality data by the National Center for Water Quality Research at Heidelberg University and the United States Geological Survey. The authors declare that they have no conflicts of interest.

References

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

Arabkhedri, M., Lai, F. S., Noor-Akma, I., & Mohamad-Roslan, M. K. (2010). An application of adaptive cluster sampling for estimating total suspended sediment load. *Hydrology Research*, *41*(1), 63–73. https://doi.org/10.2166/nh.2010.113

Arnab, R. (2017). Chapter 5 - unequal probability sampling. In R.Arnab (Ed.), *Survey sampling theory and applications* (pp. 117–166). Academic Press. https://doi.org/10.1016/b978-0-12-811848-1.00005-4

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*, 3029–3047. https://doi.org/10.1002/hyp.6147

Barbiero, A., & Mecatti, F. (2010). Bootstrap algorithms for variance estimation in πPS sampling. In P.Mantovan & P.Secchi (Eds.), *Complex data modeling and computationally intensive statistical methods*. Contributions to Statistics: Springer. https://doi.org/10.1007/978-88-470-1386-5_5

Beale, E. (1962). Some uses of computers in operational research. *Industrielle Organisation*, *31*(1), 27–28.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer. Retrieved from https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/

Cassidy, R., & Jordan, P. (2011). Limitations of instantaneous water quality sampling in surface-water catchments: Comparison with near-continuous phosphorus time-series data. *Journal of Hydrology*, *405*(1–2), 182–193. https://doi.org/10.1016/j.jhydrol.2011.05.020

Chanat, J. G., Moyer, D. L., Blomquist, J. D., Hyer, K. E., & Langland, M. J. (2016). *Application of a weighted regression model for reporting nutrient and sediment concentrations, fluxes, and trends in concentration and flux for the Chesapeake Bay Nontidal Water-Quality Monitoring Network, results through water year 2012*. U.S. Geological Survey Scientific Investigations Report 2015.5133; U.S. Geological Survey. https://doi.org/10.3133/sir20155133

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

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

D'Agostino, R. B., & Stephens, M. A. (Eds.), (1986). *Goodness-of-Fit Techniques*. Marcel Dekker.

De Vries, A., & Klavers, H. C. (1994). Riverine fluxes of pollutants: Monitoring strategy first, calculation methods second. *European Water Management*, *4*(2), 12–17.

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*: Chapman & Hall.

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

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

Ferguson, R. I. (1986b). Reply [to "Comment on 'River loads underestimated by rating curves' by R. I. Ferguson"]. *Water Resources Research*, *22*(13), 2123–2124. https://doi.org/10.1029/WR022i013p02123

Gregoire, T. G., & Valentine, H. T. (2007). *Sampling strategies for natural resources and the environment*. CRC Press.

Guy, H. P., & Norman, V. (1970). Field methods for measurement of fluvial sediment. *U.S. Geological Survey techniques of water-resources Investigations*, book 3, chap: C2; U.S. Geological Survey.

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

Hansen, M., & Hurwitz, W. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, *14*(4), 333–362. https://doi.org/10.1214/aoms/1177731356

Heidelberg University. (2020). *Heidelberg University National Center for Water Quality Research Tributary data download*, Retrieved on 7 Dec. 2020 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

Hirsch, R. M., & De Cicco, L. A. (2015). *User guide to exploration and graphics for RivEr Trends (EGRET) and dataRetrieval.R packages for hydrologic data* (version 2.0, February 2015). *U.S. Geological Survey Techniques and Methods*, *4*, 93. https://doi.org/10.3133/tm4A10

Hirsch, R. M., Moyer, D. L., & Archfield, S. A. (2010). Weighted regressions on time, discharge, and season (WRTDS), with an application to Chesapeake Bay river inputs. *Journal of the American Water Resources Association*, *46*(5), 857–880. https://doi.org/10.1111/j.1752-1688.2010.00482.x

Horowitz, A. J. (2010). The use of instrumentally collected composite samples to estimate the annual fluxes of suspended sediment and sediment associated chemical constituents. *Sediment Dynamics for a Changing Future*, *337*, 273–281.

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 Ng it right? *Environmental Science & Technology*, *47*(6), 2471–2486. https://doi.org/10.1021/es304058q

Horowitz, A. J., Clarke, R. T., & Merten, G. H. (2015). The effects of sample scheduling and sample numbers on estimates of the annual fluxes of suspended sediment in fluvial systems. *Hydrological Processes*, *29*(4), 531–543. https://doi.org/10.1002/Hyp.10172

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685. https://doi.org/10.1080/01621459.1952.10483446

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

Jones, A. S., Horsburgh, J. S., Mesner, N. O., Ryel, R. J., & Stevens, D. K. (2012). Influence of sampling frequency on estimation of annual total phosphorus and total suspended solids loads. *Journal of the American Water Resources Association*, *48*(6), 1258–1275.

Jung, H., Senf, C., Jordan, P., & Krueger, T. (2020). Benchmarking inference methods for water quality monitoring and status classification. *Environmental Monitoring and Assessment*, *192*(4), 261. https://doi.org/10.1007/s10661-020-8223-4

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

Koch, R. W., & Smillie, G. M. (1986a). 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

Koch, R. W., & Smillie, G. M. (1986b). Comment on “River loads underestimated by rating curves” by R. I. Ferguson. *Water Resources Research*, *22*(13), 2121–2122. https://doi.org/10.1029/WR022i013p02121

Lahiri, D. B. (1951). A method of sample selection providing unbiased ratio estimates. *Bulletin Institut International de Statistique*, *33*, 133–140.

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; 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.); U.S. Geological Survey. https://doi.org/10.3133/ofr20171120

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

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

Lohr, S. L. (2021). *Sampling design and analysis* (3rd ed.,). Chapman and Hall/CRC.

Luo, P., He, B., Chaffe, P. L., Nover, D., Takara, K., & Mohd Remy Rozainy, M. A. (2013). Statistical analysis and estimation of annual suspended sediments of major rivers in Japan. *Environmental Science Processes & Impacts*, *15*(5), 1052–1061. https://doi.org/10.1039/c3em30777h

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, *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

Miller, M. P., Tesoriero, A. J., Hood, K., Terziotti, S., & Wolock, D. M. (2017). Estimating discharge and nonpoint source nitrate loading to streams from three end-member pathways using high-frequency water quality data. *Water Resources Research*, *53*(12), 10201–10216. https://doi.org/10.1002/2017wr021654

Oelsner, G. P., Sprague, L. A., Murphy, J. C., Zuellig, R. E., Johnson, H. M., Ryberg, K. R., et al. (2017). *Water-quality trends in the Nation’s rivers and streams, 1972-2012—Data preparation, statistical methods, and trend results* (ver. 2.0 (p. 136). U.S. Geological Survey Scientific Investigations. https://doi.org/10.3133/sir20175006

Owen, A. B. (1992). Empirical likelihood and small samples. In C.Page & R.LePage (Eds.), *Computing Science and statistics* (pp. 79–88). Springer. https://doi.org/10.1007/978-1-4612-2856-1_10

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. (1968). *Sampling theory*: McGraw-Hill Book Company.

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, Chap. A5*: U.S. Geological Survey. https://doi.org/10.3133/tm4a5

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

Tada, A. (2021). *Daily water quality data sets of the nine U.S. Watersheds*. Zenodo. https://doi.org/10.5281/zenodo.5154774

Tada, A., & Tanakamaru, H. (2021). Unbiased estimates and confidence intervals for riverine loads. *Water Resources Research*, *57*(3). https://doi.org/10.1029/2020wr028170

Tanner, M. A. (2006). *Tools for Statistical Inference* (3rd ed.). Springer. 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. (1988). Measuring sediment yields of storms using PSALT. *Sediment Budgets. IAHS Publ. no.*, *174*, 315–323.

Thomas, R. B. (1989). Piecewise SALT sampling for estimating suspended sediment yields. In *Pacific Southwest Forest and Range Experiment Station*: U.S. Forest Service. General Technical Report, PSW-114. Retrieved from https://www.fs.usda.gov/treesearch/pubs/8636

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. Geological Survey. (2009). *Monitoring large rivers in the National stream quality accounting Network (NASQAN). U.S. Geological Survey information Sheet, 2pp*. Retrieved from https://water.usgs.gov/nawqa/nasqan.information.09252008.pdf

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

U.S. Geological Survey. (2021). *Methods and glossary (network information) in tracking water quality in U.S. streams and rivers, USGS national water quality network data, water-quality loads, and trends*. Retrieved from https://nrtwq.usgs.gov/nwqn/#/TECH

Verhoff, F. H., Yaksich, S. M., & Melfi, D. A. (1980). River nutrient and chemical-transport estimation. *Journal of the Environmental Engineering Division*, *106*(ee3), 591–608. https://doi.org/10.1061/jeegav.0001047

Walling, D. E., & Webb, B. W. (1981). The reliability of suspended sediment load data. *Erosion and Sediment Transport Measurement*, *133*, 177–194.

Water Framework Directive. (2003). *Common implementation strategy for the water framework directive* (*2000/60/EC*). *Guidance Document*, *7*. Retrieved from https://ec.europa.eu/environment/water/water-framework/facts_figures/guidance_docs_en.htm

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), 1–2721. Retrieved from https://library.csbe-scgab.ca/docs/journal/49/c0625.pdf

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