Water Resources Management (2019) 33:4303–4318
https://doi.org/10.1007/s11269-019-02369-7

The Value of Intensive Sampling—A Comparison of Fluvial Loads

Saurav Kumar[1] · Adil Godrej[2] · Harold Post[2] · Karl Berger[3]

Received: 13 June 2019 / Accepted: 26 September 2019 /
Published online: 4 November 2019

Abstract

Most long-term sampling regimes are calendar based, collecting one or two samples per month regardless of the stream conditions. Loads estimated with calendar-based sampling are often used for expensive water quality mitigation measures. In this paper, we have tested the differences between the calendar-based and extensive sampling methods for two watersheds of different sizes, and three parameters—total nitrogen, total phosphorus, and total suspended solids. Based on the results obtained and the costs associated with the remediation, a simple decision-making framework is proposed for watershed managers to decide on the applicability of a calendar-based sampling method. Direct loads (DL) were computed using a method based on an intensive sampling of flow and other water quality parameters. Weighted regression loads (WL) were estimated using the WRTDS model designed for modified calendar-based sampling. The results suggest that for trend analysis and planning on a larger scale, long-term loads obtained from a modified calendar-based sampling regime may be used as a reasonable substitute for loads obtained from intensive sampling. However, for purposes where accurate daily loads are needed (e.g., water quality model calibration) WL may not be an effective substitute for DL. Finally, we recommend that the costs of control measures should be assessed when deciding on a sampling regime.

Keywords Nutrient loads · Sampling methods · Regression methods · WRTDS · Remediation costs

✉ Saurav Kumar
saurav@tamu.edu

[1] Department of Biological and Agricultural Engineering and Texas A&M AgriLife Research El Paso, 1380 A&M Circle, El Paso, TX 79927, USA

[2] Department of Civil and Environmental Engineering, Virginia Tech, 9408 Prince William St, Manassas, VA 20110, USA

[3] Metropolitan Washington Council of Governments, 777 North Capitol Street NE, Suite 300, Washington, DC 20002, USA

1 Introduction

Knowing fluvial loads and concentrations of pollutants is essential to establishing the state of a waterbody and its tributary watershed, and assessing trends in water quality. Load estimates of several parameters, such as Total Nitrogen (TN), Total Phosphorous (TP), and Total Suspended Sediments (TSS), are used to establish budgets for nonpoint and point pollution sources and design remediation/mitigation strategies. The regulatory framework in the United States used to establish total maximum daily loads (TMDLs) for impaired waterbodies was created in response to the Clean Water Act and its modifications (FWPCA 2002). The framework is heavily dependent on using accurate and reliable computation of fluvial loading to control degradation and restore designated use(s). Computation of loads, particularly for regulatory compliance, is still not standardized and may never be because it is based on data, and resources allocated to collect data vary widely.

There are several sampling methods, with strengths and weaknesses, that may be employed for estimating loads. Methods that rely on extensive sampling produce reliable load estimates but are resource-intensive to undertake, whereas methods that do not need extensive sampling often are not very accurate, particularly at shorter time scales. Decision-makers may try to optimize the *desired accuracy* in load estimation with the costs involved in obtaining these estimates for their region (optimization decision). The problem often comes when trying to express the *desired accuracy* as a cost that may be compared with the costs of obtaining accurate loads. One potential way to estimate the value/cost of the *desired accuracy* is to use the cost of management measures for pollution abatement that may be driven by the optimization decision.

With the success that the United States has enjoyed in controlling the majority of point sources, further reduction of fluvial loads for regulatory purposes requires nonpoint or diffuse pollution abatement measures that are often expensive to control. Typically, the costs of diffuse pollution controls are very high, and control measures themselves are often unreliable. If best management practices (BMPs) with limited control ability and those that are practically infeasible in producing results (e.g., pet waste management education, control of illicit discharges, and reduction in urban growth) are excluded from consideration, the costs per pound (over the lifecycle including capital and operational costs) of controlling non-point pollution of TN, TP, and TSS are between $151–$14,449, $1851–$70,342, and $4—-$69, respectively, for some areas of the Chesapeake Bay watershed region (CWP 2013). The wide ranges are reflective of the type of BMP, efficiency in controlling the pollutant, and other local conditions such as soil type and land value. Further, it may be noted there are theoretical reasons to believe that the cost of reducing impairment will be of a convex shape, where the costs per unit reduction decrease first with the economies of scale and then increase with the diminishing marginal returns on resources invested in impairment reduction (Wainger 2012). It may be reasonable to believe that in most scenarios, where the easier, cheaper methods have already been implemented, the economies of scale have been exhausted and further reductions will require significantly higher costs per unit reduction of impairment.

1.1 Fluvial Loads Sampling and Computation Methods

The most accurate estimates for loads may be obtained using continuous measurements of flow rates and in-stream concentrations of the water quality parameters of interest. Near-continuous recording (every 15 min to hourly) of flow measurement may be done using a variety of techniques, and some parameters such as temperature, nitrate-nitrogen, and dissolved oxygen, may also be measured in a near-continuous fashion. However, the measurement of many parameters requires laboratory analysis. This requirement practically rules out long-term high-frequency water quality measurement. Nevertheless, with judiciously frequent water quality measurements, excellent estimates of fluvial loads may be made (He et al. 2018; Johnes 2007; Kronvang and Bruhn 1996; Kumar et al. 2013; Moyer et al. 2012; Park and Engel 2015; Stenback et al. 2011). For example, the concentration for any parameter in a stream typically does not change much during non-storm baseflow periods. Thus, very good to excellent load estimates may be obtained for non-storm periods using weekly or bi-weekly water quality sampling. During storm events, when the concentrations of the parameters of interest (and, consequently, load) may be expected to vary rapidly, compositing methods of load estimation that yield an event mean concentration (EMC) for the storm may be employed to get accurate storm-event loads.

In resource-constrained scenarios, reasonably frequent composite storm or base flow sampling are not feasible, and regression methods are often used for estimating constituent concentrations. Regression-based methods estimate the constituent concentration by relating flow and other readily measurable parameters to the concentration of the constituent of interest (He et al. 2018; Kumar et al. 2013). Typically, regression-based methods require calendar-based (e.g., monthly) sampling as the measured concentration data are only used for calibration of a regression equation. The putative trade-off of this method is the reliability of the loads estimated at a lower cost.

There is considerable evidence suggesting that in smaller watersheds calendar-based sampling methods do not perform adequately, and even in larger watersheds significant differences were found in long-term studies (Horowitz et al. 2015; Kumar et al. 2013; Robertson and Roerish 1999; Stenback et al. 2011). To improve the efficiency of the regression-based fluvial load using calendar-based sampling, different sampling methods, such as hydrological-based sampling, storm chasing, sampling in the rising or falling limb of the hydrograph, and adaptive cluster sampling have been used with varying degrees of success (Arabkhedri et al. 2010; Horowitz et al. 2015; Robertson and Roerish 1999; Sadeghi et al. 2008; Sadeghi and Saeidi 2010).

The United States Geological Survey (USGS) has developed a modified weighted regression method–Weighted Regression on Time, Discharge, and Season (WRTDS)–to address some of the issues with regression-based fluvial load estimation schemes (Hirsch et al. 2010). The WRTDS method was shown to perform well in several scenarios (Beck and Hagy 2015; Lee et al. 2016; Sprague et al. 2011; Zhang et al. 2016). The good performance of WRTDS when used with modified calendar-based sampling and the prevalence of the method, particularly after being adopted by the USGS, were the motivations for using WRTDS for this study.

1.2 Decision Making Framework for Monitoring

A comprehensive framework for developing a monitoring program (Fig. 1) relies on linked relations between several components including the natural system under observation, the objective of the monitoring program, sampling scheme, field collection and analysis methods, and the cost-effectiveness (Maher et al. 1994). The natural system under investigation and the desired objective is used to derive an observed indicator. The cost-effectiveness and sampling scheme can then be optimized. Broad monitoring design plans, such as the US TMDL Effectiveness Monitoring Plan and the EU Water Framework Directive direction on monitoring plans (Allan et al. 2006) are often too broad and may not be directly applicable to choosing between the various options available for processes in Fig. 1. In this study, we develop and assess a limited simplified framework relating the cost-effectiveness and performance of the modeling method to choose a sampling method for resource-constrained monitoring operations.

2 Study Area

This study utilized data collected for two stations, ST30 and PR01, marked in Fig. 2, which also shows their drainage areas. Data for both stations were obtained from the Occoquan Watershed Monitoring Laboratory (OWML). ST30 on Broad Run in northern Virginia drains an area of about 2.29×10^2 km^2, and PR01 on the Potomac River drains an area of about 2.9×10^4 km^2. The much smaller ST30 watershed is relatively uniform in elevation and slopes, whereas the PR01 watershed area spans four states and includes parts of the Appalachian Mountains and has a much higher variation in elevation and slopes.

3 Methods

3.1 Load Computation

The two methods of load computation utilized in this study are:

1) Direct Method, which used the OWML dataset of weekly/bi-weekly water quality data for the three parameters of interest during non-storm flows and EMC for storm events, along with near-continuous (15 min to hourly) flow measurements. The direct method represents the best model for estimating load with extensive sampling schemes.
2) WRTDS Method, which is based on a weighted-regression technique described by Hirsch et al. (2010). In this study, for the WRTDS method, we used the same OWML weekly/bi-weekly water quality data during non-storm flows but used discrete samples that were also

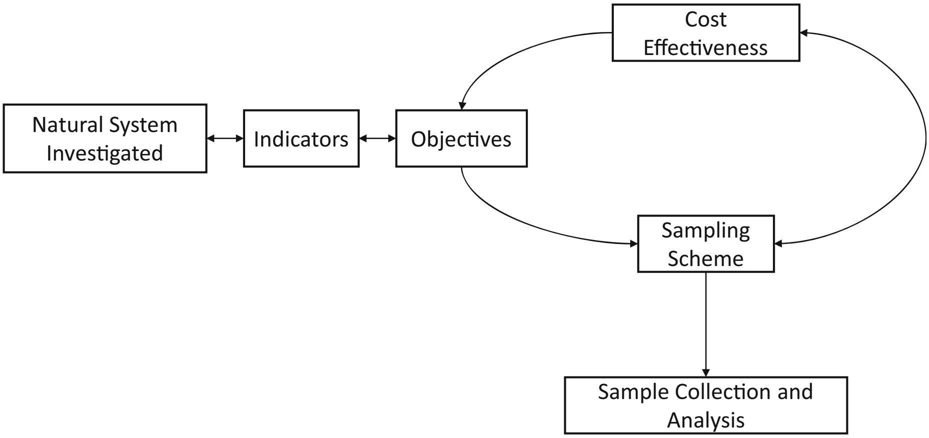

Fig. 1 A simple framework for designing a sampling program adapted from Maher et al. (1994)

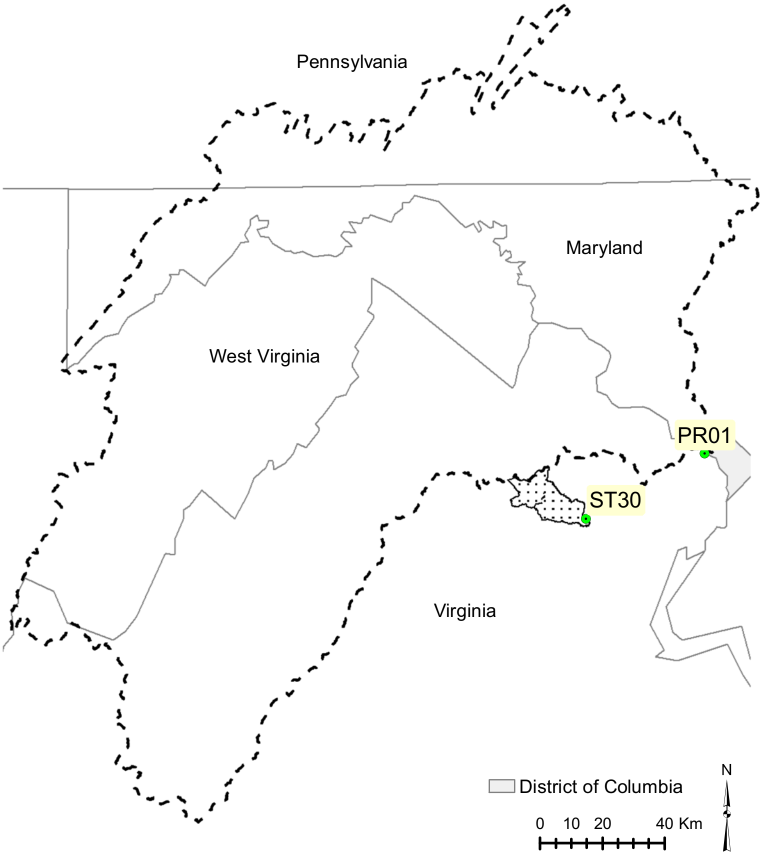

Fig. 2 Location of the two stations used in this study. The watershed for PR01 (on Potomac River) station spans four states and the District of Columbia. The Little Falls Dam is a short distance upstream from the PR01 station. ST30 station (on Broad Run) is the smaller watershed abutting the Potomac River watershed but not a part of it

taken during storm events, along with the daily average flow. The WRTDS method is the representative method used for estimating loads with modified calendar-based sampling schemes.

3.1.1 Direct Method

The direct method of fluvial load computation is an extension of the first principle of load calculation. In this method, a water quality concentration reading is assigned for every recorded flow reading. For non-storm flows, regular periodic discrete samples measurements (weekly, bi-weekly) are used to interpolate concentrations at every data point where the flow is recorded (usually hourly during non-storm flows). For storm events, the flow-composite EMC value is assigned to all the recorded storm time flows; see Kumar et al. (2013) for more details on the composite sampling method employed. Loads for the desired period are then computed by aggregation. In this paper, we have considered this load (Direct Load or DL) to be the reference and all the bias is computed with relation to this reference.

3.1.2 WRTDS Method

The WRTDS uses a five-parameter equation (Eq. 1) to estimate concentration based on flow and time. This approach is different from most other regression methods as the parameters of Equation 1 $\left(\hat{\beta}_0 \text{ to } \hat{\beta}_4\right)$, instead of being fixed, are estimated for every combination of Q (daily average flow) and t (time) where concentration has to be computed. Thus, a unique set of coefficients is estimated for every combination of Q and t in the period of record. The advantage of this approach is the ability to be unbiased and still estimate a wider class of regression surfaces than other parametric functions used for most other regression methods. The data set used to compute $\hat{\beta}_0$ *to* $\hat{\beta}_4$ is weighted based on the distance from Q and t at the estimation point. Weight (w) is computed for three different distances: a) time, b) season, and c) discharge using the "tri-cube weight function" (Equation 5). Net weight is taken as the product of these three weights. A 7-year half-window width is used for trend distance, 0.5 (decimal time) half-window width is used for seasonal distance, and 2.0 (ln[Q]) half window width for weight computation. These half-window widths (h) are similar to what was used and found to be optimum by USGS for non-tidal load computation in the Chesapeake Bay watershed (Hirsch et al. 2010; Moyer et al. 2012). All load computations for the WRTDS methods were performed by using the R Software Exploration and Graphics for RivEr Time-series (EGRET) package developed by USGS [https://github.com/USGS-R/EGRET/wiki, Access Date: 02/14/2019]. A much more detailed explanation about the WRTDS method may be obtained from the software webpage [https://github.com/USGS-R/EGRET/wiki, Access Date: 02 /14/2019] and USGS publication (Hirsch and De Cicco 2015).

$$\ln(c) = \hat{\beta}_0 + \hat{\beta}_1 \ln(Q) + \hat{\beta}_2(t) + \hat{\beta}_3 \sin(2\pi t) + \hat{\beta}_4 \cos(2\pi t) + \varepsilon \tag{1}$$

where

c is concentration $\frac{mg}{l}$
Q is observed daily flow, $\frac{m^3}{s}$
t is the decimal time, years
$\hat{\beta}_0$ to $\hat{\beta}_4$ are regression coefficient estimates
ε is the unexplained variation

$$w = \begin{cases} \left(1 - \left(\frac{d}{h}\right)^3\right) & \text{if } |d| \leq h \\ 0 & \text{if } |d| > h \end{cases} \tag{2}$$

where

w is the weight
d is the distance from estimation point to data point
h is the half-window width

Calibration and Performance of WRTDS To assess the performance of the WRTDS model the coefficient of determination (r^2) based on observed and predicted concentrations was computed. The computation was on daily concentrations, not loads. An r^2 value of greater than 0.6 is considered satisfactory for water quality modeling (Donigian 2002). Note that several other detailed methods of assessing performance based on residual analysis and others are available in the EGRET package. Only r^2 was used in this study and it was hard to assess the performance using other, often graphical, methods, in a simplified decision-making scenario.

3.1.3 Load Comparisons

Load flux at two stations was computed by two methods, WRTDS (WL) and Direct Load Method (DL), for three parameters (TN, TP, and TSS) at three averaging timescales (annual, monthly, and daily). A total of thirty-six (36) fluvial load time series were computed (2 stations, 2 methods, 3 parameters, and 3 averaging times) for the period from 1989 to 2003. This fifteen-year period was chosen because during this period OWML was collecting discrete samples during storm events (necessary for calibrating WRTDS) at ST30. In 2004, discrete storm sampling was discontinued.

Matched-pair comparisons were made to establish the difference in daily, monthly, and annual load fluxes computed by the two approaches. Because normality of the differences between pairs of flux could not be established (even after log transformation), non-parametric matched-pair signed-rank test (using R *wilcox.test*) was performed, comparing log-transformed DL with WL.

Matched-pair signed-rank test, where the difference between the two datasets is tested for the null hypothesis of zero, was used to identify the statistically significant (α =0.05) difference. An unbiased magnitude-of-difference (δ) between loads was calculated with the "Hodges-Lehmann Estimator" as suggested by Helsel and Hirsch (2002) to estimate the difference for the fifteen-year period. The Hodges-Lehmann Estimator (Equation 3) represents the difference between two populations and is computed as the median of all possible pairwise differences. For daily time series, the difference data (for matched-pair sign rank test) were found to be serially correlated. The Autoregressive (AR1) model was found to be suitable for the daily difference data using the autocorrelation function and partial autocorrelation function plots (not shown). Testing was thus performed on 'pre-whitened' data as discussed in von Storch (1995) for AR(1) removal. Pre-whitening was performed using the formula in Equation 4 and the residual time series was used for testing.

$$\delta = \mathrm{median}\left\{ \left(X_i - X_j\right); i = 1, \ldots, n; j = 1, \ldots, n; j \neq i \right\} \tag{3}$$

where

δ is the Hodges-Lehmann Estimator;
Xs are the differences;

$$X'_t = X_t - r_1 X_{t-1} \tag{4}$$

where

X_t is the daily difference at time t;

X'_t is the residual pre-whitened data used for testing;
r_1 is the lag1 sample serial correlation.

To estimate flow independent concentration trends, the Kendall-Theil robust slope was computed on residuals obtained after using a LOWESS function to explain variations in the average concentration with the observed average flows for the period. These trends were computed for all cases. All statistical tests were performed using well-established R (*wilcox.test*) and Python (*theilslopes* in *scipy.stats*) libraries. Further data partitioning based on flow (Low, Medium, and High) was done to analyze the difference in loads computed by DL and WL for different categorical flows. For the data partitioning, *low flows* were defined as periods (daily, monthly or annual) where the average flows (averaged over the time period of interest: annual, monthly, or daily) were from 0 to 25 percentile of observed flows, *medium flows* were from 26 to 74 percentile, and *high flows* were from 75 to 100 percentile.

4 Results

To calibrate the WRTDS model and compute WL, an average of 43 and 52 samples per year for ST30 and PR01, respectively, were used. In addition to the discrete samples used to calibrate WRTDS, additional flow composite sampling was performed for the computation of DL during the study period. Based on the r^2 all the models (except TP and TSS at ST30) indicated acceptable results (>0.6) on a daily timescale. Models for TP and TSS with r^2 of 0.55 and 0.50 were still used as WRTDS estimated loads and represented the best method to compute loads from a modified calendar-based sampling regime where only daily flows are available. It may be noted that WRTDS is not recommended for small flashy watersheds where the flows may vary substantially within a day (Hirsch and De Cicco 2015).

Figures 3 and 4 show the differences between the two stations for three timescales. For both ST30 and PR01, the interquartile range (represented by the width of the box plot) of the difference between the two load computation methods is lower for the annual timescale followed by monthly, and then daily. The spread for the difference quantified as the interquartile range for TN is smaller than that of TP and TSS for both stations. The majority of the difference [$\Delta = \ln(\text{DL}) - \ln(\text{WL})$] is negative for daily time scales at both stations, suggesting that WL load > DL load. For monthly and annual, the median difference is closer to zero.

Table 1 shows the significance of the differences and the multiplicative retransformed magnitude-of-difference (δ). Differences were computed on a natural log scale and hence are multiplicative, not additive. A statistically significant δ is observed for all parameters on a daily time scale, except for TN at PR01. At the monthly timescale, δ for TP and TSS at PR01 are significant. At the annual timescale, δ for TN at PR01 and TSS at ST30 were found to be significantly different from zero.

Table 2 shows the slopes for the flow-independent concentration trendline. None of the slopes for the trendline for annual data by either method was found to be significantly different from zero. Significant trends were observed from monthly and daily flow-independent data, except for TN at PR01 on a monthly timescale. Significance and the direction of the trendline, represented by the sign of the slope, computed by data from both methods were similar for all time scales and parameters. The magnitude of slopes calculated varied widely for TSS but was largely similar (within the 95% confidence bound) for TN and TP.

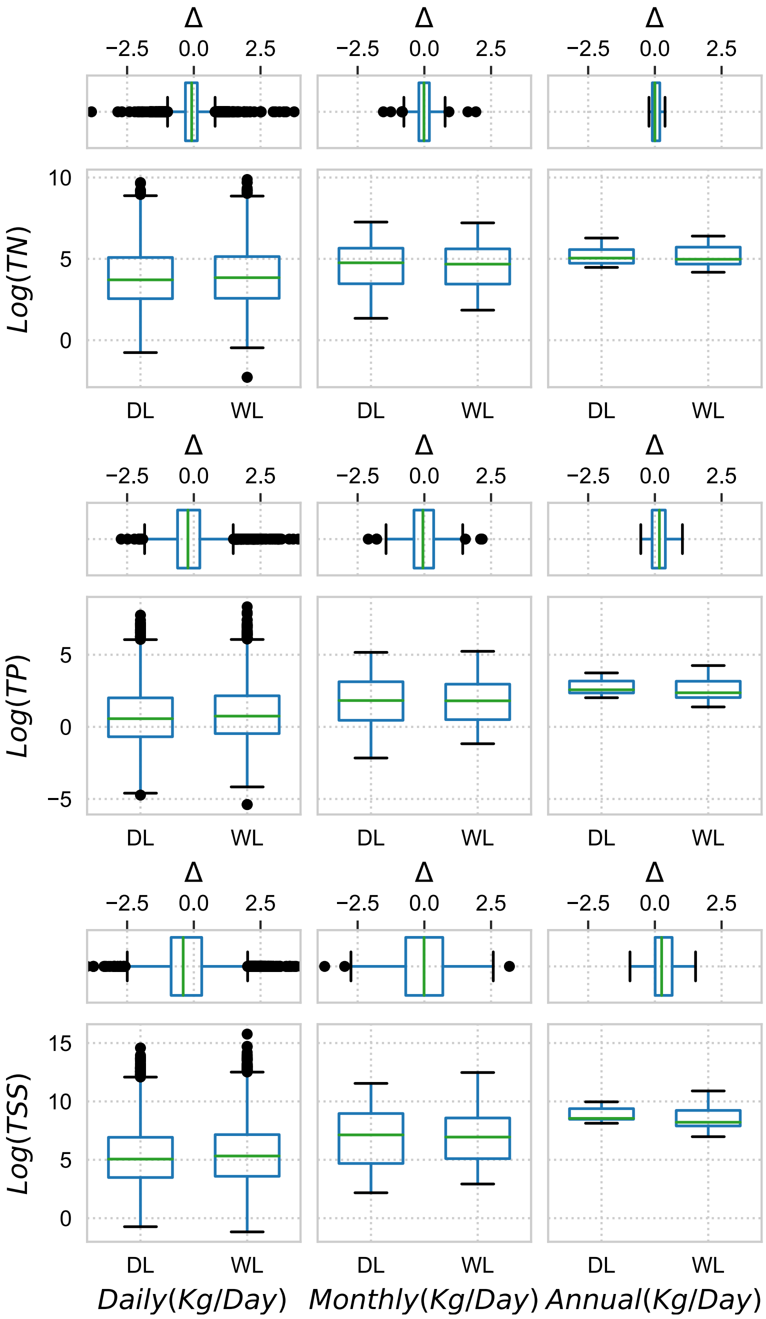

Fig. 3 Load flux and the log differences observed at ST30. The vertical boxplots show the log-transformed flux in Kg/day, and the horizontal boxplots show the difference (Δ) computed as ln(DL)-ln(WL)

Analysis of the difference in the methods after flow partitioning (high, medium, and low) suggests that these differences seem to vary with the flow partition, although no pattern is evident. As was observed in Figs. 3 and 4, the difference in prediction of TN is less than TP

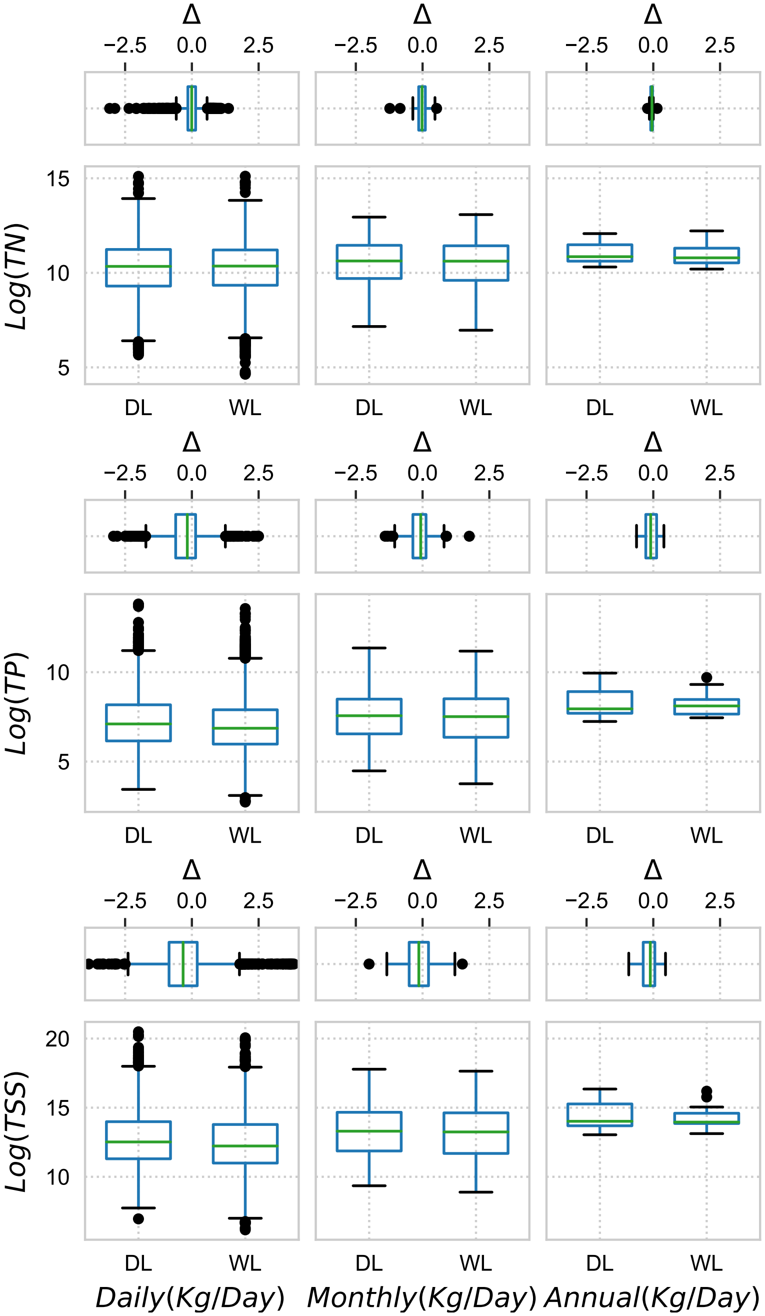

Fig. 4 Load flux and the log difference observed at PR01. The vertical boxplots show log-transformed flux computed in Kg/day, and the horizontal boxplots show the difference (Δ) computed as ln(DL)-ln(WL)

and TSS for all computed scenarios. With DL as the reference, WL seems to under-predict loads for all three parameters on an annual timescale at ST30 during low flows; the WL prediction at other times are more closely aligned with DL. At PR01, on annual timescales for

Table 1 The statistical significance of the difference between the two methods of flux computation and the magnitude of difference. Note that the sample size for annual, monthly, and daily comparison are 15, 180 (15 × 12), and 5477, respectively

	PR01			ST30		
	TN	TP	TSS	TN	TP	TSS
	***p*-values for matched-pair Sign-Rank test**					
Annual	**0.004***	0.229	0.188	0.421	0.188	**0.048***
Monthly	0.159	**< 0.001***	**0.002***	0.563	0.492	0.832
Daily	0.190	**< 0.001***	**< 0.001***	**< 0.001***	**< 0.001***	**< 0.001***
	Multiplicative magnitude of difference δ (DL = WL × δ)					
Annual	**0.9***	0.9	0.9	1	1.1	**1.4***
Monthly	1	**0.9***	**0.9***	1	1	1
Daily	1	**0.8***	**0.7***	**0.9***	**0.8***	**0.7***
	Percent difference (100 × [WL-DL]/DL) **					
Annual	**11***	11	11	0	-9	**−29***
Monthly	0	**11***	**11***	0	0	0
Daily	0	**25***	**43***	**11***	**25***	**43***

***Bold* * values are statistically significant difference at $\alpha = 0.05$**

***% difference computed as 100 × (WL-DL)/DL = 100× (1- δ)/ δ*

TP and TSS, there is wide under-prediction and over-prediction by WL at Low and High flows, respectively. TN at PR01 on annual timescale seems to be invariant to the flow partitions.

5 Discussion

In natural systems, it is reasonable to expect that larger/sudden variations in concentration and flows will occur during storm events. The sampling method used for DL captures all storm event loads and is likely to yield more representative or true loads. Thus, all load bias discussion here is referenced to DL.

Table 2 Slopes (changes in mg/L per year) of the trend line along with the 95% confidence bound. Note that a negative slope shows a decline of the parameter and a positive slope an increase. The magnitude of the slope represents the average change over the study period for the parameter of interest

		PR01		ST30	
		DL	WL	DL	WL
Annual	**TN**	−0.02 (−0.04,0.01)	−0.03 (−0.04,0.01)	0.00 (−0.02,0.02)	0.01 (−0.00,0.01)
	TP	−0.001 (−0.005,0.004)	−0.001 (−0.006,0.007)	0.000 (−0.003,0.003)	0.002 (−0.000,0.003)
	TSS	−0.6 (−3.1,1.7)	−1.0 (−4.3,3.4)	0.5 (−1.4,2.9)	1.7 (−0.7,4.0)
Monthly	**TN**	−0.01 (−0.03,0.00)	−0.01 (−0.02,0.00)	**0.02*(0.01,0.03)**	**0.02*(0.01,0.02)**
	TP	**0.002*(0.000,0.004)**	**0.002*(0.001,0.004)**	**0.002*(0.001,0.003)**	**0.002*(0.001,0.003)**
	TSS	**0.7*(0.0,1.5)**	**0.9*(0.3,1.7)**	**0.6*(0.2,1.1)**	**0.5*(0.3,0.8)**
Daily	**TN**	**−0.01*(−0.02,−0.01)**	**-0.01*(−0.01,-0.01)**	**0.03*(0.02,0.03)**	**0.02*(0.02,0.02)**
	TP	**0.002*(0.002,0.002)**	**0.002*(0.001,0.002)**	**0.001*(0.001,0.002)**	**0.002*(0.002,0.002)**
	TSS	**0.4*(0.3,0.4)**	**0.5*(0.4,0.6)**	**0.1*(0.1,0.1)**	**0.2*(0.2,0.2)**

***Bold* * values are statistically significant difference at $\alpha = 0.05$**

The performance of the modified-calendar based sampling regime represented by WL for PR01 seems to be good ($r^2 > 0.6$). The low r^2 for TP and TSS at ST30 does question the applicability of a modified calendar-based sampling load computation for ST30. ST30 with shorter flow events (less than a day) associated with storms may be classified as a small watershed. For small watersheds, a model based on daily flows may not capture all variations. It is important to recognize that loads are a product of flow and concentration and much higher flow values (measured more reliably) may mask this poor performance of the modified calendar-based sampling. Increasing the load averaging timescale from daily to monthly or annually may also improve performance.

The values of δ close to 1 for both ST30 and PR01 on annual and monthly periods (Table 1), except for TSS at ST30 on annual time averaging scale, suggest a lack of any systematic bias in the load computation with WL when compared to DL as the reference. For TSS at ST30 on an annual timescale, WL overestimates loads by about 29% over the study period. On daily timescales, however, evidence of systematic bias exists: 5 out of 6 observations being statistically significant, and 4 out of 6 showing more than a 10% difference. Some uncertainty is expected in sample collection and wet chemistry testing of various constituents. For a small watersheds Harmel et al. (2006) have quantified the "typical-minimum" net uncertainty for TN, TP, and TSS load measurement to be in the order of 11%, 8%, and 7%, respectively, and this can serve as a good reference for evaluating the bias in the loads estimated by WL (typical maximums are defined as 70%, 110%, 53%, and typical averages as 29, 30%, 18%). Using typical-minimum, a conservative assumption, most of the long-term loads on annual and monthly timescales are within the limits of expected errors in DL (Table 2). Daily loads are generally more variable and outside the typical minimum values.

Flow-independent trends produced by the two methods are similar in direction and statistical significance. The magnitude of the trend slopes are similar (not exactly the same, but within the confidence bounds) except for TSS at ST30. The observation of low overall δ and similar trends indicate that for long-term planning, where variation in one period is not very important, modified calendar-based sampling may be a good substitute for both the large PR01 and the small ST30 watersheds. The statistical formulation used for computing WL also allows for direct computation of a flow-normalized trend (different from flow-independent trends discussed here) that may make the method more attractive for long-term trend analysis.

The WL method may not represent reality in cases where variations in each time period may be important, such as calibrating water quality models at a daily time-step for source allocation of loads. Regression methods and WRTDS have been used in the Chesapeake Bay watershed to calibrate parts of the Bay Model (Easton et al. 2017). For a watershed model calibrated with WL at a daily time step, it is reasonable to expect that the model will try to account for the wide daily discrepancy (from DL) which will result in inadequate system representation with the parameters adjusted to match the WL. The performance of WRTDS for smaller ST30 watershed (low r^2 for 2 out of 3 parameters) strongly suggests that WRTDS should be avoided for smaller watersheds. Hirsch and De Cicco (2015) have suggested that WRTDS with a daily time step may not be appropriate for small flashy watersheds, defined as "where discharge at the stream gage commonly changes by an order of magnitude or more within a given day."

As discussed in the introduction, strategic sampling during wet or dry weather has been suggested to improve the performance of regression methods. Our results based on flow partition suggest that it is not easy to determine when to sample more (flows where load computation errors are frequent). In the two watersheds studied, the difference between WL and DL varied with the parameter, flow, and station in a manner such that no discernable correlation could be observed. It is important to note that the number of samples per year to drive the WRTDS, on average 43 for ST30 and 52 for PR01, in this study is high for a typical resource-constrained watershed monitoring program. Even for the extensively monitored nine large rivers in the Chesapeake Bay watershed, less than an average of 20 samples per year are collected, with huge variation among parameters. WRTDS protocol recommends at least 20 samples (12 calendars and 8 stormflow). The number of samples collected per year from 1985 to 2010 for some rivers are as low as 6 for suspended sediments and as high as 41 for TN and TP at the Potomac River (Moyer et al. 2012). No attempt has been made in this study to assess the impact of the reduction in samples. With a reduction in samples, the results of this study may not hold and likely will be worse for regression-based methods. Shorter-term parallel sampling, sampling intensively with EMC for all storms and grab samples for some storms that will allow computation of both WL and DL, as suggested by Kumar et al. (2013), may be used to identify and correct the systematic biases in regression-based methods.

The choice of sampling method is often based on available resources. From an economic perspective, the costs associated with the construction and upkeep of nonpoint BMPs may be used as a surrogate for the benefits obtained from the investment in sampling. In a watershed where low-cost nonpoint BMP options such as street sweeping are used, extensive sampling needed to estimate DL may not be necessary. As the costs of control measures per pound of pollutant controlled rise, investment in extensive sampling becomes more feasible either to control cost if the regression methods are overpredicting or to confirm the loads and ensure attainment of load targets. Even the slight statistically significant over-prediction by the WRTDS method for TN at PR01 on an annual timescale ($\delta = 0.9$) may predict 2400 Kg/day to be remediated. In a typical TMDL allocation scenario, this load would be distributed among point and non-point sources. Given that the median urban stormwater TN remediation cost is about 3160 $/Kg (James River Watershed), even a fraction of the 2400 Kg/day that may be assigned to the urban region will result in an extremely high excess cost, in the order of millions of dollars per day.

Without analysis such as the one presented in the study, there is no good way to understand the biases in loads predicted by modified calendar-based sampling, and the cost of remediation if WL overpredict is very high. It may be argued that from a policymaker's perspective there is no scenario where modified calendar-based sampling should be used without intensive sampling comparison. In cases where WL loads are underpredicted, the situation may lead to non-attainment of designated use: not a desirable and often expensive option. Nevertheless, a simple framework for decision making (Fig. 5) on whether or not to use modified calendar-based sampling may rely on the performance of the regression equation, type of control measures available, and the estimated bias in the use of the regression equation. If the performance of the regression equation is 'acceptable' and 'inexpensive' control measures are available, the application of WL may be justified. The 'acceptable' measure can be based on literature (e.g., $r^2 > 0.6$). 'Inexpensive' control measures are more subjective and may depend on the location being analyzed. If either of these conditions is not met, a rigorous sampling regime of 'parallel sampling' may be undertaken, where sampling will allow for the computation of both DL and WL. Kumar et al. (2013) have estimated that 12–72 months parallel sampling may be required for computing an accurate estimate of the load based on the parameter and size of the watershed. Finally, based on bias estimation and comparison with typical errors expected in load computation, a decision on whether to use modified calendar-based sampling may be made. This framework can be applied for any new or existing

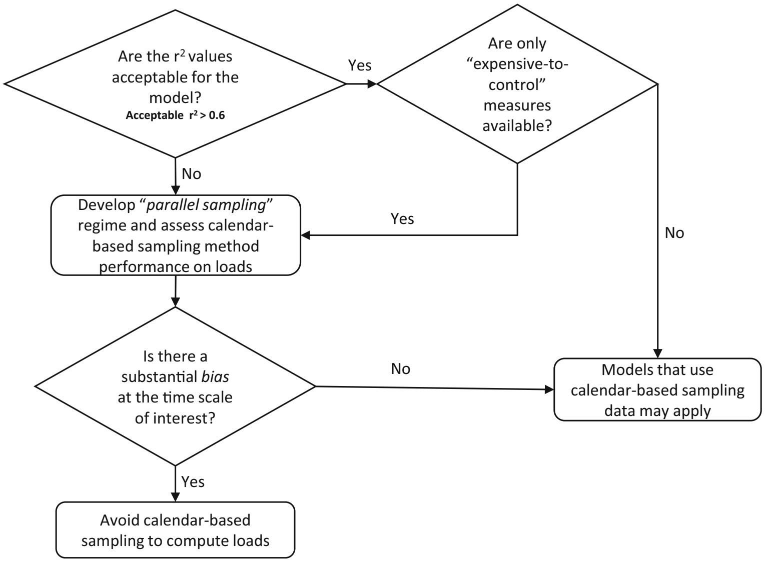
Are the r^2 values acceptable for the model?
Acceptable $r^2 > 0.6$
Yes
Are only "expensive-to-control" measures available?
No
Develop "*parallel sampling*" regime and assess calendar-based sampling method performance on loads
Yes
No
Is there a substantial *bias* at the time scale of interest?
No
Models that use calendar-based sampling data may apply
Yes
Avoid calendar-based sampling to compute loads

Fig. 5 A simple decision-making framework to assess the applicability of the models that use modified calendar-based sampling data

monitoring program. Especially where non-point control measures may be required in the short-term. The strength of this framework is the simplicity and ability to assess the bias objectively. However, the framework may often recommend parallel sampling that may require monitoring for years before a long-term method may be employed.

6 Conclusion

In this study, for the watersheds analyzed, a statistical lack of similarity between the loads computed by modified calendar-based sampling and the loads computed via extensive sampling could be shown at the annual and monthly timescales for 4 comparisons: PR01 (TN annual, TP and TSS monthly), and ST30 (TSS annual). There is not enough evidence to reject similarity for eight other scenarios: PR01 (TN monthly, TP and TSS annual), and ST30 (TN and TP monthly and annual, and TSS monthly). Thus, if the goal of a monitoring program is estimating long-term annual mean loadings that may be used for TMDL (or similar) allocation scenario either DL or WL may work for some parameters. For the daily loads, 5 out of 6 comparisons show a difference, with TN at PR01 being the exception. Thus, for a goal of estimating short-term daily loading that may be used for calibrating water quality model and captures storm loads WL is unlikely to provide reliable estimates. Overall, these results indicate that there may be an agreement between loads computed by WL and DL for exactly 50% of all cases (9 out of 18).

The flow-independent trends showed a better correlation compared to fluvial loads, with similar magnitudes for all statistically significant slopes and the same direction of the trend at the two stations for all three averaging times. Thus if the goal for the monitoring program is to estimate long-term trends useful for assessing progress towards achieving long-term targets either WL or DL may be used.

Given that the remediations often needed to control the non-point loads are very expensive, we have argued that parallel sampling, which allows for load computation by both DL and WL, should be conducted for some period. Using a modified calendar-based sampling method directly without that information may not be advisable. Some degree of parallel sampling will allow watershed managers to perform comparisons like the one presented in this paper and assess periodic and long-term bias. In the absence of any prior information, an exhaustive sampling method to enable the framework discussed should be employed.

Acknowledgments The authors would like to acknowledge the stakeholders of the Metropolitan Washington Council of Governments for their continued (since 1982) support of the Potomac River (PR01) monitoring program and (since 1973) the Occoquan Watershed Monitoring Program (ST30). We will also like to acknowledge the work by Dr. Thomas Grizzard who was instrumental in developing this study. Unfortunately, he passed away before this manuscript could be written and is not included as an author as per the journal policy.

Compliance with Ethical Standards

Conflict of Interest The authors are aware of no conflict of interest.

References

Allan IJ et al (2006) Strategic monitoring for the European water framework directive. TrAC Trends Anal Chem 25:704–715. https://doi.org/10.1016/j.trac.2006.05.009

Arabkhedri M, Lai F, Noor-Akma I, Mohamad-Roslan M (2010) An application of adaptive cluster sampling for estimating total suspended sediment load. Hydrol Res 41:63–73

Beck MW, Hagy JD (2015) Adaptation of a weighted regression approach to evaluate water quality trends in an estuary. Environmental Modeling & Assessment 20:637–655. https://doi.org/10.1007/s10666-015-9452-8

Donigian A (2002) Watershed model calibration and validation: The HSPF experience Proc WEF 2002:44–73

Easton Z et al. (2017) Scientific and Technical Advisory Committee Review of the Phase 6 Chesapeake Bay Watershed Model. STAC Publication Number 17–007. Edgewater, MD, USA

FWPCA (2002) Fedral Water Pollution Control Act vol 33 U.S. Code § 1251. United States Senate and House of Representatives, Washington DC

Harmel R, Cooper R, Slade R, Haney R, Arnold J (2006) Cumulative uncertainty in measured streamflow and water quality data for small watersheds. Trans ASABE 49:13

He Y, Gui Z, Su C, Chen X, Chen D, Lin K, Bai X (2018) Response of sediment load to hydrological change in the upstream part of the Lancang-Mekong River over the past 50 years. Water 10:888. https://doi.org/10.3390/w10070888

Helsel D, Hirsch R (2002) Statistical Methods in Water Resources. Techniques of Water-Resources Investigations of the United States Geological Survey, Book 4, Hydrologic Analysis and Interpretation, chapter A3. U.S. Geological Survey, Reston, VA, USA

Hirsch RM, De Cicco LA (2015) User guide to Exploration and Graphics for RivEr Trends (EGRET) and dataRetrieval: R packages for hydrologic data, Version 1.0: Originally posted October 8, 2014; Version 2.0: February 5, 2015 edn., Reston, VA, USA. doi:https://doi.org/10.3133/tm4A10

Hirsch RM, Moyer DL, Archfield SA (2010) Weighted regressions on time, discharge, and season (WRTDS), with an application to Chesapeake Bay River inputs. J Am Water Resour Assoc 46:857–880. https://doi.org/10.1111/j.1752-1688.2010.00482.x

Horowitz AJ, Clarke RT, Merten GH (2015) The effects of sample scheduling and sample numbers on estimates of the annual fluxes of suspended sediment in fluvial systems. Hydrol Process 29:531–543. https://doi.org/10.1002/hyp.10172

Johnes P (2007) Uncertainties in annual riverine phosphorus load estimation: impact of load estimation methodology, sampling frequency, baseflow index and catchment population density. J Hydrol 332:241–258

Kronvang B, Bruhn A (1996) Choice of sampling strategy and estimation method for calculating nitrogen and phosphorus transport in small lowland streams. Hydrol Process 10:1483–1501

Kumar S, Godrej AN, Grizzard TJ (2013) Watershed size effects on applicability of regression-based methods for fluvial loads estimation. Water Resour Res 49:7698–7710. https://doi.org/10.1002/2013WR013704

Lee CJ, Hirsch RM, Schwarz GE, Holtschlag DJ, Preston SD, Crawford CG, Vecchia AV (2016) An evaluation of methods for estimating decadal stream loads. J Hydrol 542:185–203. https://doi.org/10.1016/j.jhydrol.2016.08.059

Maher WA, Cullen PW, Norris RH (1994) Framework for designing sampling programs. Environ Monit Assess 30:139–162. https://doi.org/10.1007/BF00545619

Moyer DL, Hirsch RM, Hyer KE (2012) Comparison of Two Regression-Based Approaches for Determining Nutrient and Sediment Fluxes and Trends in the Chesapeake Bay Watershed, U.S. Geological Survey Scientific Investigations Report 2012–5244. U.S. Geological Survey Richmond, VA, USA

Park YS, Engel BA (2015) Analysis for regression model behavior by sampling strategy for annual pollutant load estimation. J Environ Qual 44:1843–1851

Robertson DM, Roerish ED (1999) Influence of various water quality sampling strategies on load estimates for small streams. Water Resour Res 35:3747–3759

Sadeghi SHR, Saeidi P (2010) Reliability of sediment rating curves for a deciduous forest watershed in Iran. Hydrol Sci J 55:821–831

Sadeghi SHR et al (2008) Development, evaluation and interpretation of sediment rating curves for a Japanese small mountainous reforested watershed. Geoderma 144:198–211. https://doi.org/10.1016/j.geoderma.2007.11.008

Sprague LA, Hirsch RM, Aulenbach BT (2011) Nitrate in the Mississippi River and its tributaries, 1980 to 2008: are we making Progress? Environ Sci Technol 45:7209–7216. https://doi.org/10.1021/es201221s

Stenback GA, Crumpton WG, Schilling KE, Helmers MJ (2011) Rating curve estimation of nutrient loads in Iowa rivers. J Hydrol 396:158–169. https://doi.org/10.1016/j.jhydrol.2010.11.006

von Storch H (1995) Misuses of statistical analysis in climate research. In: von Storch H, Navarra A (eds) Analysis of climate variability. Springer, New York, pp 11–26

Wainger LA (2012) Opportunities for reducing Total maximum daily load (TMDL) compliance costs: lessons from the Chesapeake Bay. Environ Sci Technol 46:9256–9265. https://doi.org/10.1021/es300540k

Zhang Q, Harman CJ, Ball WP (2016) An improved method for interpretation of riverine concentration-discharge relationships indicates long-term shifts in reservoir sediment trapping. Geophys Res Lett 43:10, 215–210,224. https://doi.org/10.1002/2016GL069945

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.