JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

AMERICAN WATER RESOURCES ASSOCIATION

ESTIMATION OF CATCHMENT NUTRIENT LOADS IN NEW ZEALAND USING MONTHLY WATER QUALITY MONITORING DATA[1]

T.H. Snelder, R.W. McDowell, and C.E. Fraser[2]

ABSTRACT: Causes of variation between loads estimated using alternative calculation methods and their repeatability were investigated using 20 years of daily flow and monthly concentration samples for 77 rivers in New Zealand. Loads of dissolved and total nitrogen and phosphorus were calculated using the Ratio, L5, and L7 methods. Estimates of loads and their precision associated with short-term records of 5, 10, and 15 years were simulated by subsampling. The representativeness of the short-term loads was quantified as the standard deviation of the 20 realizations. The L7 method generally produced more realistic loads with the highest precision and representativeness. Differences between load estimates were shown to be associated with poor agreement between the data and the underlying model. The best method was shown to depend on the match between the model and functional and distributional characteristics of the data, rather than on the contaminant. Short-term load estimates poorly represented the long-term load estimate, and deviations frequently exceeded estimated imprecision. The results highlight there is no single preferred load calculation method, the inadvisability of "unsupervised" load estimation and the importance of inspecting concentration-flow, unit load-flow plots and regression residuals. Regulatory authorities should be aware that the precision of loads estimated from monthly data are likely to be "optimistic" with respect to the actual repeatability of load estimates.

(KEY TERMS: streamflow; watershed management; nutrients; nonpoint source pollution; statistics.)

Snelder, T.H., R.W. McDowell, and C.E. Fraser, 2016. Estimation of Catchment Nutrient Loads in New Zealand Using Monthly Water Quality Monitoring Data. *Journal of the American Water Resources Association* (JAWRA) 1-21. DOI: 10.1111/1752-1688.12492

INTRODUCTION

Management of diffuse sources of nutrients discharging into aquatic environments is an important issue facing most countries with intensive agriculture (Daily *et al.*, 1997; McDowell *et al.*, 2015). Policy responses that aim to decrease the load of nutrients lost from land to water include restricting nutrient inputs to land such as restricting nitrogen fertilizer or slurry inputs in nitrate vulnerable zones (Environment Agency, 2013) and prohibiting certain land use practices such as application of manure in winter (Iowa State Legislature, 2010). Irrespective of the approach, policies and actions must be linked to quantified nutrient discharges and measurable environmental impact at the catchment scale (Meals, 1996).

Agriculture has intensified in New Zealand over the last 20 years, in particular, dairy farming, resulting in localized eutrophication of some water bodies due to increased nutrient loads (Parliamentary

[1]Paper No. JAWRA-16-0120-P of the *Journal of the American Water Resources Association* (JAWRA). Received May 5, 2016; accepted October 19, 2016.

[2]Director (Snelder), LWP Limited, 145c Colombo Street, Beckenham, Christchurch 8023, New Zealand; Principal Scientist (McDowell), Land and Environment Group, AgResearch Limited, Mosgiel 9053, New Zealand; and Senior Engineer (Fraser), Aqualinc Research Limited, Christchurch 8053, New Zealand (E-Mail/Snelder: ton@lwp.nz).

Commissioner for the Environment, 2013). In response to this, recent government legislation in New Zealand, the National Policy Statement for Freshwater Management (NPS) (Ministry for the Environment, 2014), requires provincial regulatory authorities (regional councils) to clarify freshwater objectives and to promulgate measures, such as catchment nutrient load limits, to ensure that the objectives can be achieved. In response to the NPS, regional councils are increasingly estimating nutrient discharge limits for catchments, and devising management strategies and monitoring programs to ensure that these limits are met (Roygard *et al.*, 2012). Accurate estimates of average annual catchment nutrient loads and their precision are needed to calibrate models that underlie these strategies and to quantify the degree of change in a load that can be considered statistically significant. Key questions that must be addressed by regulatory authorities are: (1) What is the mean annual catchment load? and (2) Has the load changed and does it currently exceed the catchment load limit? These are not trivial questions and must consider the availability of suitable data, differences in estimates associated with alternative load calculation methods, and the uncertainty associated with the load estimates.

Catchment nutrient loads are generally calculated from a combination of discrete low-frequency measurements of concentration (*e.g.*, monthly) and flow, which is often measured at higher frequencies (*e.g.*, daily) (Defew *et al.*, 2013; Dolan *et al.*, 1981; Hirsch, 2014; Preston *et al.*, 1989; Quilbé *et al.*, 2006). Numerous methods are used to calculate loads from flow records and water quality data. Two common categories of load calculation methods are: (1) ratio methods and (2) rating methods (*e.g.*, Cohn, 2005; Defew *et al.*, 2013). Ratio methods are based on the assumption that the ratio of the mean of the instantaneous load to the mean instantaneous flow associated with individual samples is representative of the ratio of the long-term mean load to the mean flow (Beale, 1962). Rating methods define an empirical relationship between concentration and flow and use this to generate an estimate of concentration for every observation of flow. Allowance for various factors, such as flow magnitude and season can also be incorporated in both these categories of methods (Dolan *et al.*, 1981; Preston *et al.*, 1989; Quilbé *et al.*, 2006).

Many studies have considered the uncertainty of catchment nutrient load estimates calculated using different methods and sampling frequencies by comparing calculated loads with a "true load" that has been calculated from concentrations sampled at high temporal resolution but over a relatively short period (<1 year) (*e.g.*, Defew *et al.*, 2013; Johnes, 2007; Phillips *et al.*, 1999; Preston *et al.*, 1989; Quilbé *et al.*, 2006; Robertson and Roerish, 1999). Uncertainty comprises two components: bias, or average difference between the estimated and true load; and imprecision, which describes the variance of the estimates. In general, studies have found that one load calculation method will often provide the least uncertain estimate for a specific site and contaminant; but that no method is consistently superior. The uncertainty of the methods has been shown to depend on many factors including the frequency of sampling (Defew *et al.*, 2013; Robertson and Roerish, 1999), duration of the sampling period (Littlewood *et al.*, 1998), the size of the watershed (Phillips *et al.*, 1999), the distributional characteristics of the contaminants (Preston *et al.*, 1989; Young *et al.*, 1988), the strength and form of the flow-concentration relationship (Cohn, 2005; Preston *et al.*, 1989; Richards and Holloway, 1987), the catchment sources of the contaminants (Johnes, 2007), and the characteristics of the flow regime (Johnes, 2007; Preston *et al.*, 1989). The most appropriate method is therefore widely regarded as being dependent on the characteristics of the site and contaminant (Kronvang and Bruhn, 1996; Quilbé *et al.*, 2006).

While these previous studies provide useful insights into the shortcomings of low-frequency monitoring data for estimating loads, regulatory authorities often find themselves in a position where they need to calculate catchment loads from limited data. Water quality monitoring on a monthly basis is common practice for monitoring water quality state and trends in New Zealand and in other countries. For the majority of catchments, the data available to estimate nutrient loads are a combination of these monthly concentrations and continuous flow records. The periods over which these data are available are variable and often relatively short (*e.g.*, less than a decade; Larned *et al.*, 2004). Loads calculated from these data using different methods can vary considerably. However, there is a lack of objective methods to assess which method provides the most robust estimates. It has also been shown that distributional and functional characteristics of concentration and discharge can differ by large amounts between years (Young *et al.*, 1988). Thus, load estimates made from monthly concentration data for short time periods may poorly represent loads estimated for another time period, and their associated precision may under estimate the true repeatability of the estimate. However, these issues have not been examined, despite their importance in the calibration of models and determination of minimum detectable changes in loads.

To address these issues, this study aimed to: (1) quantify the magnitude of differences in nutrient

loads estimated using different methods; (2) quantify the precision and representativeness of nutrient loads calculated using different sampling periods and compare these results across sample period durations, methods, and nutrient species; and (3) assess whether there are underlying distributional and functional relationships that influence these differences. To achieve these goals, we used three methods to calculate loads at multiple sites using monthly concentration and continuous (daily) flow for four nutrient species: Dissolved Reactive Phosphorus (DRP), nitrate-nitrogen (NO_3-N), Total Nitrogen (TN), and Total Phosphorus (TP). The "true" load could not be estimated because we were restricted to monthly concentration data at all sites; however, the data were sufficient to address the study aims.

METHODS

Nutrient Concentration and Flow Data

We used data from the National Rivers Water Quality Network (NRWQN) (Davies-Colley *et al.*, 2011), which comprises 77 sites located on 48 of New Zealand's rivers (Figure 1). The NRWQN sites broadly represent variation in the catchments of main-stem rivers across New Zealand and exhibit a wide range of flow magnitude, flow variability, water quality, and catchment character including area and proportion of catchment occupied by pastoral land use (Table 1), which is strongly related to nutrient concentrations (Larned *et al.*, 2004).

Since 1989, a range of water quality variables have been sampled punctually at monthly intervals and flows have been monitored at 15-min intervals (Davies-Colley *et al.*, 2011; Smith and McBride, 1990). In this study, we used observations of flow, DRP, NO_3-N, TN, and TP for the time period 1991-2010 (20 years). Samples of dissolved nutrient analyses were filtered (Whatman GF/F; Maidstone, UK) in the laboratory before analysis (Smith and Maasdam, 1994). NO_3-N + NO_2-N, and DRP concentrations were measured with a QuikChem 8000 flow-injection analyzer (Lachat Instruments, Milwaukee, Wisconsin). NO_2-N concentrations were <1% of NO_3-N concentrations, and NO_3 + NO_2-N is referred to hereafter as NO_3-N. Concentrations of TN and TP were measured with the flow-injection analyzer after persulfate digestion (Smith and Maasdam, 1994).

Samples for which the concentrations were at or below the detection limit (0.5, 1, 1, and 2 mg/m^3 for DRP, NO_3-N, TN, and TP, respectively) were set to half the detection limit. This treatment of censored values was considered acceptable because only TP samples were ever below the detection limit and on only 17 occasions (0.1%). We used mean daily flows and assumed that the measured concentrations were representative of the flow-weighted daily mean concentrations (Quilbé *et al.*, 2006; Robertson and Roerish, 1999).

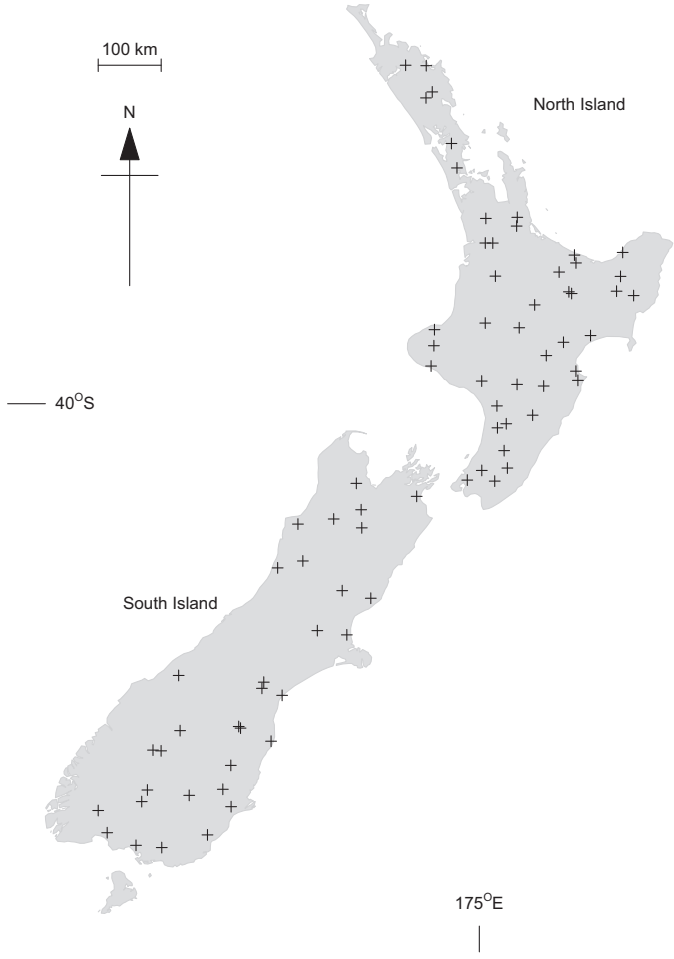

FIGURE 1. Location of the 77 Study Sites.

For some of the analyses that follow, it was important that there were no trends in the concentration data. However, nutrient concentration data for the NRWQN sites are subject to monotonic trends, both positive and negative (Scarsbrook *et al.*, 2003). We therefore de-trended concentration data by regressing the log (base 10)-transformed site concentrations against the sample dates. When there was a statistically significant ($p < 0.05$) relationship between date and concentration, we adjusted the concentrations by adding the back-transformed residuals of the regression models to the long-term mean concentration. The adjusted data represented a time series with no monotonic trend and a mean equal to the original mean.

TABLE 1. Summary of Characteristics of Catchments, Flow Regimes, and Nutrient Concentrations at the 77 Sites.

Characteristics	Description	Units	Mean	Min	Max
CatA	Catchment area	km^2	3,204	14	20,504
Qmean	Mean daily flow	m^3/s	108	1	578
Pasture	Proportion of catchment occupied by pasture land cover	%	22	0	86
γ_q	Skewness of the site flow distribution		3.4	−0.5	14.8
CV_q	Coefficient of variation in the annual mean flow at the sites		0.22	0.08	0.43
Median; DRP	Site median concentrations of nutrient species	mg/m^3	9	0.5	71
Median; NO_3-N			233	0.7	1,862
Median; TN			369	40	2,148
Median; TP			24	3	120
γ_c; DRP	Skewness of the distribution of site concentrations of		2.3	−0.5	9.5
γ_c; NO_3-N	the four nutrient species		1.7	−0.3	9.5
γ_c; TN			2.7	−0.2	10.3
γ_c; TP			7.1	0.9	14.8

Note: DRP, Dissolved Reactive Phosphorus; NO_3-N, nitrate-nitrogen; TN, Total Nitrogen; TP, Total Phosphorus.

Load Calculation Methods

If flow and concentration observations were available for each day, the total load would be the summation of the daily flows multiplied by their corresponding concentrations:

$$L = \frac{K}{A_c} \sum_{j=1}^{N} C_j Q_j \quad (1)$$

where L: mean annual load expressed as an export coefficient (kg/yr/ha), A_c: catchment area, ha, K: units conversion factor (31.6 kg s/mg/year), C_j: concentration for each day in period of record (mg/m^3), Q_j: daily mean flow for each day in period of record (m^3/s^{-1}), and N: number of days in period of record.

In this summation, the individual products represent unit loads. Because concentration data are generally only available for infrequent days, unit loads can only be calculated for these days. However, flow is generally observed continuously and there is generally a relationship between concentration and flow or unit load. Alternative load calculation methods exploit these relationships to improve the accuracy of loads estimated from infrequent concentration observations. To explore these relationships, we plotted flow *vs.* concentration and unit loads for all sites and nutrient species to visually assess the relative strength of these relationships within our dataset.

For each site, we calculated the mean annual load for each of the four nutrient species using three commonly used and recommended methods, which we refer to as the Ratio method, the seven-parameter (L7) rating method, and the five-parameter (L5) rating method. We expressed all nutrient loads as annual export coefficients (*i.e.*, kg/year/ha) by dividing the annual load (kg/year) by the catchment area (ha).

Ratio Method. The Ratio method calculates the average unit load, based on the days when concentrations were observed (Dolan *et al.*, 1981; Quilbé *et al.*, 2006). This average unit load is then adjusted by the ratio of the mean flow for all days to the mean flow on days when concentrations were observed. The adjusted average unit load is multiplied by the number of days to determine the total load. The adjustment of the average unit load made by the Ratio method assumes two conditions: (1) that there is a positive linear relationship between the unit loads and flows, which passes through the origin (Cochran, 1977); and (2) that the variance of load is proportional to flow (Cochran, 1977; Preston *et al.*, 1989). Because flow and load are almost always correlated, this ratio is biased (Cochran, 1977). Beale (1962) developed a correction term that adjusts for this bias by accounting for the covariance between the unit loads and flows. The ratio method, with Beale's bias correction, is given by:

$$L = \frac{K}{A_c} \bar{Q} \frac{\bar{l}}{\bar{q}} \left(\frac{1 + \frac{1}{n} \times \frac{\mathrm{cov}(l_i, q_i)}{\bar{l}\bar{q}}}{1 + \frac{1}{n} \times \frac{\mathrm{var}(q_i)}{\bar{q}^2}} \right) \quad (2)$$

where q_i: daily mean flow on days with concentration samples (m^3/s), $l_i = c_i q_i$ is the daily loads on the sample days (kg), $\bar{l} = \sum_{i=1}^{n} \frac{c_i q_i}{n}$ is the mean load on the days with concentration samples (kg), and $\bar{q} = \sum_{i=1}^{n} \frac{q_i}{n}$ is the mean daily flow for the entire period (m^3/s).

L7 Model. Rating methods derive a relationship between the sampled nutrient concentrations (C_i) and flow (Q_i), which is then used to estimate concentration for each day of the entire sampling period (Cohn *et al.*, 1989). The estimated concentrations and daily flows are combined to estimate unit loads for each day and these are summed for the entire period.

Two regression model approaches to defining rating curves of (Cohn *et al.*, 1989, 1992) and (Cohn, 2005) are commonly used in the United States and New

Zealand (Alexander *et al.*, 2002). The regression models relate the log of concentration to the sum of three explanatory variables: discharge, time, and season. The L7 model is based on seven fitted parameters given by:

$$\ln\left(\widehat{C}_i\right) = \beta_1 + \beta_2\left[\ln(q_i) - \overline{(\ln(q))}\right] + \beta_3\left[\ln(q_i) - \overline{(\ln(q))}\right]^2 + \beta_4\left(t_i - \overline{T}\right) + \beta_5\left(t_i - \overline{T}\right)^2 + \beta_6\sin(2\pi t_i) + \beta_7\cos(2\pi t_i) \quad (3)$$

where, $\beta_{1,2,\ldots7}$: regression coefficients, t_i: time in decimal years, $\overline{T}$: mean value of time in decimal years, $\overline{(\ln(q))}$: mean of the natural log of discharge on the sampled days, and $\widehat{C}_i$: estimated concentration at time i.

The coefficients are estimated from the sample data by linear regression, and when the resulting fitted model is significant ($p < 0.05$), it is then used to estimate the concentration on each day in the sample period. The resulting estimates of $\ln(\widehat{C}_i)$ are back-transformed (by exponentiation) to concentration units. Because the models are fitted to the log transformed concentrations, the back-transformed predictions are corrected for retransformation bias. We used the smearing estimate (Duan, 1983) as a correction factor (S):

$$S = \frac{1}{n}\sum_{i=1}^{n} e^{\hat{\varepsilon}_i} \quad (4)$$

where, $\hat{\varepsilon}$ are the residuals of the regression models. The smearing estimate assumes that the residuals are homoskedastic and therefore the correction factor is applicable over the full range of the predictions.

The total load is then calculated by combining the flow and estimated concentration time series:

$$L = \frac{K}{A_c}\left(\sum_{j=1}^{N} S\frac{\widehat{C}_j Q_j}{N}\right) \quad (5)$$

If the fitted model is not significant, $(\widehat{C}_j)$ is replaced by the mean concentration and S is unity.

L5 Model. The L5 model is the same as L7 model except that two quadratic terms are eliminated:

$$\ln\left(\widehat{C}_i\right) = \beta_1 + \beta_2(\ln(q_i)) + \beta_3(t_i) + \beta_4\sin(2\pi t_i) + \beta_5\cos(2\pi t_i) \quad (6)$$

The five parameters are estimated and loads are calculated in the same manner as the L7 model.

Precision of Load Estimates

The statistical precision of a sample statistic, in this study, the mean annual load, is the amount by which it can be expected to fluctuate from the population parameter it is estimating due to sample error. In this study, the precision represents the repeatability of the estimated load if it was reestimated using the same method under the same conditions. Precision is characterized by the standard deviation of the sample statistic, commonly referred to as the standard error. We evaluated the standard error of each load estimate by bootstrap resampling (Efron, 1981). For each load estimate, we constructed 100 resamples of the concentration data (of equal size to the observed dataset), each of which was obtained by random sampling with replacement from the original dataset. Using each of these datasets, we recalculated the site load and estimated the precision (SE) as the standard deviation of the 100 estimates.

Comparing Magnitude of Loads Estimated Using Different Methods

We calculated the loads and their standard errors for each site, method and nutrient species combination using the full 20 years of data. The magnitudes of these load estimates were compared by plotting the cumulative frequency distributions of site export coefficients. Differences in loads calculated using different methods were quantified as:

$$\Delta L_{AB} = \frac{L_A - L_B}{L_A} \times 100 \quad (7)$$

where, ΔL_{AB}: percent difference in site loads calculated using methods A and B (%), L_A: site export coefficient for load calculation method A (kg/year/ha), and L_B: site export coefficient for load calculation method B (kg/year/ha).

We assessed the proportion of significantly different load estimates for the three pairs of contrasting methods using the loads estimated from the bootstrap samples. We considered that loads calculated using contrasting methods were significantly different if the 95% confidence interval for the differences between corresponding bootstrap samples did not contain zero.

Explanation of Differences in Site Loads Calculated Using Different Methods

We quantified the suitability of the regression models that defined the rating curves that underlie the L5 and L7 methods based on (1) the adjusted

coefficients of determination (r^2 value) and (2) an examination of the residuals of these regression models. The adjusted-r^2 only increases with additional model terms if the increase in explained variation is greater than can be expected by chance. The adjusted-r^2 therefore provided a comparable measure of suitability of the L5 and L7 models. We multiplied the residuals by their associated flow and used the mean flow-weighted residual value (ε) as a second measure of the suitability of the rating curve. Flow weighting recognizes that the effect of deviations between the observations and the rating curve on the estimated load is proportionate to the flow at which these deviations occur (Equation 1). We standardized ε by first expressing the residuals in concentration units by back-transformation (by exponentiation) and then dividing both the flow and concentration values by dividing by their respective mean values.

We hypothesized that the variation in loads calculated using different methods is associated with the validity of the models underlying each method. We tested this for the rating methods using Kendall's rank correlation between the differences in site loads calculated using different pairs of methods (ΔL_{AB}) and ε and r^2. Kendall's coefficient (τ), measures the association between two variables based on their rank correlation: the similarity of the orderings of the data when ranked. We were not able to define a similar quantitative measure for the Ratio method but were able to subjectively assess the degree to which the assumptions of the method were reasonable from the plots of flow versus the unit loads.

Variation in Load Estimates with Time Period and Sample Duration

We simulated the effect of estimating loads using short-term datasets by subsampling the 20-year dataset. For each site, nutrient, and method, we calculated loads using subsampled short-term periods of the (detrended) time series having three durations (5, 10, and 15 years). For each duration, we made 20 separate time series realizations by taking periods comprising consecutive complete calendar years that started at the beginning of each calendar year of record (*i.e.*, 1991, 1992, ... 2010). For periods that extended beyond the last year of record (2010), we substituted years by beginning the time series again from 1991.

For each realization, we calculated the load and its standard error using the three methods. For each method, this resulted in 20 loads and associated standard error estimates for each combination of site, duration, and nutrient. We used the mean of the 20 standard errors (SE_P) to represent the characteristic precision associated with a site, duration, and nutrient. We used the standard deviation of the 20 load estimates to quantify how representative the individual short-term load estimates were of the characteristic mean annual load. This "representativeness" indicates the variation in the sample statistic (*i.e.*, the mean annual load) and can be regarded as the standard error associated with differing time period (SE_R). The standard deviation was calculated using the log (base 10)-transformed loads because the distribution of the 20 loads were strongly right skewed; the $\log_{10}$ loads were approximately symmetric and normally distributed.

To enable comparison of the characteristic precision (SE_P) and representativeness (SE_R), we expressed both as relative 95% confidence intervals. The relative 95% confidence interval for the characteristic precision was calculated by multiplying SE_P by the appropriate z-score (1.96) and dividing by the load estimate. The relative 95% confidence interval for representativeness was calculated as:

$$CI_P = \frac{L_P - 10^{[\log_{10}(L_P) \pm 1.96 \times SE_R]}}{L_P} \times 100 \qquad (8)$$

where, L_P: mean site export coefficient for a specific site, nutrient, method, and sample duration (kg/year/ha).

Between-Site Differences in Load Precision and Representativeness

We hypothesized that differences in site precision (SE_P) and representativeness (SE_R) is associated with the influence of the distributional and functional characteristics of concentration and flow on the calculation method. To explore this hypothesis, we examined the strength of the association of SE_P and SE_R with five explanatory variables for each nutrient species, method, and sample period duration using Kendall's rank correlation coefficient (τ).

Two explanatory variables characterized variation in the flow regimes across the sites: the coefficient of variation in the annual mean flow (CV_q) and the skewness of the flow distribution (γ_q). CV_q was computed as the ratio of the standard deviation of annual mean flow to the mean of annual mean flows. Skewness was characterized by Pearson's moment coefficient of skewness (γ), a unit-less index that characterizes the asymmetry of the distribution (Zar, 1999). The coefficient of skewness is positive when the right tail is longer and the center of mass of the distribution is concentrated to the left. The sample skewness was estimated from the flow data as:

$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{2/3}} \quad (9)$$

where x_i are individual values in a sample of n values.

The skewness of the concentration distribution (γ_c) was computed from the concentration samples (Equation 9). For the L5 and L7 methods, the suitability of the concentration-flow relationship was represented using the adjusted-r^2 of the regression models and the mean flow-weighted residuals. Because precision (SE_P) and representativeness (SE_R) are measures of variation and not difference, we used the absolute difference of mean flow-weighted residuals ($|\varepsilon|$). All analyses were undertaken using R (R Core Team, 2016).

RESULTS

Concentration-Flow Relationships

At most sites, the majority of concentration and unit load samples were associated with low to median flows, and high flows were poorly represented (Figure 2). The form of the relationships between concentration and flow differed between nutrient species and between sites. In general, concentrations increased with flow (Figure 2); but at many sites, there was evidence that these relationships were curvilinear in log-log space. The rate of change in concentration with flow could plateau or decrease at high flow, particularly for NO_3-N, or could increase at high flows, particularly for TP. The quadratic relationship between concentration and flow, which is included in the L7 rating curve, was able to represent these curvilinear relationships, whereas the linear regression models poorly represented relationships between flow and concentration at high flow at some sites (Figure 2).

The fitted L5 and L7 models were significant for 97% of the site and variable combinations. The insignificant models included DRP at six sites, and TN and NO_3-N at two sites each. Differences in the validity of the concentration-flow rating curves were reflected in the values of the flow-weighted residuals (ε). For the L5 model, the mean of these values was negative for NO_3-N, indicating that the modeled values were generally larger than the observations at high flows, and were positive for the other variables (Table 2). The mean of site values of the mean flow-weighted residual values was generally smaller for the L7 model, indicating that this model better represented relationships between flow and concentration at high flow.

The adjusted-r^2 values of the regression models underlying the L5 and L7 models varied considerably (Table 2). The mean of the r^2 values were highest for TP and lowest for DRP (Table 2). The adjusted-r^2 were only weakly correlated with the mean flow-weighted residual values (Table 2).

There was considerable variation in the distributions of the concentration data (Table 1, Figure 2). In general, the distributions of DRP and NO_3-N were less skewed than the TN distributions, and TP was the most skewed (Table 1). The distributions were generally right skewed (γ_c was positive), but were left skewed at a few sites (Table 1).

The relationship between unit loads and flow varied between sites and nutrient species (Figure 2). Unit loads increased with flow, but at many sites, there was evidence that these relationships were not linear, particularly for TP (Figure 2).

Estimated Loads

Over the 77 sites, export coefficients calculated from the 20-year dataset varied over one and a half orders of magnitude for TN, two for DRP, three orders of magnitude for NO_3-N, and over four for TP (Figure 3). Export coefficients estimated for NO_3-N at 16 sites with the L5 method were larger than the export coefficients for TN load by large margins (Figure 3). Three of the NO_3-N export coefficients estimated using the L5 method exceeded 20 kg/ha/year. The L7 method produced export coefficients for TP that exceeded 10 kg/ha/year for five sites.

Precision

The precision of site loads estimated using the 20-year dataset (SE), varied from close to zero to over 100% (Figure 4). The precision of site loads estimated using the short-term datasets decreased (*i.e.*, SE_P increased) with decrease in the sample period duration (Figure 4).

The lowest precision (largest SE and SE_P) values for most sites, nutrients, and sample period durations were associated with the L5 method. The precision of the L5 method for DRP, NO_3-N, and TN had median values of between 12% and 19% for the 5-year duration decreasing to approximately 5% for the 20-year duration (Figure 4). The median of site values for precision for the L7 and Ratio methods were generally similar for DRP, NO_3-N, and TN being approximately 10% for the 5-year duration and decreasing to approximately 5% for the 20-year duration (Figure 4).

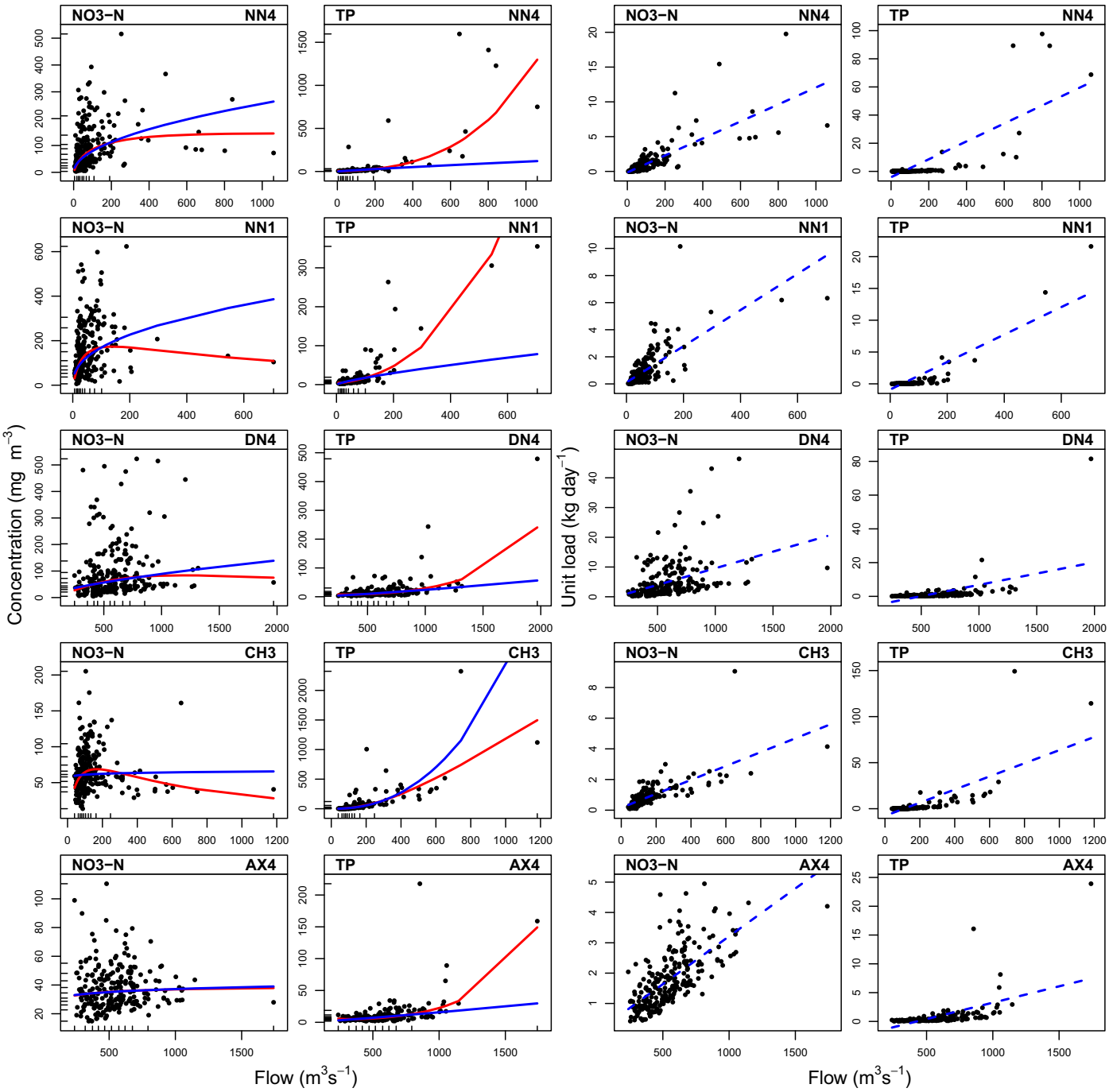

FIGURE 2. Examples of Concentration *vs*. Flow and Unit (daily) Loads *vs*. Flow for Nitrate and Total Phosphorus at Five Sites. Linear and quadratic relationships between concentration and flow (blue and red lines respectively) are shown, which were fitted to the log-transformed flow and concentration data. Linear relationships between the unit loads and flow are shown (dashed blue lines). The rug plots on the *x* and *y*-axis of the concentration-flow relationships indicate the distributions of the flow and concentration data with each line representing 10% of the data.

The lowest precision was associated with TP, for which the median site precision for the 5-year duration was 23% and 29% for the L7 and Ratio methods, respectively, and 64% for the L5 method. The median site precision for TP reduced to between 13% and 29% for the 20-year duration (Figure 4).

Differences in Load Magnitudes between Methods

The methods that produced the largest loads varied between sites (Figure 5). Differences between the L5 and Ratio methods were generally less than 30% and exceeded 30% for only 4 and 2 sites for DRP and

TABLE 2. Distributions of the Adjusted-r^2 and the Mean Flow-Weighted Residuals (ε) of the Regression Models Underlying the L5 and L7 Methods across the 77 Sites and the Correlation of the r^2 and ε Values. These variables were not applicable to the Ratio method.

		Adjusted-r^2			Mean Flow-Weighted Residuals (ε)			
Method	**Nutrient Species**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**	**Correlation of r^2 and ε**
L5	DRP	0.30	0.04	0.57	0.09	−1.69	0.80	−0.40
	NO_3-N	0.48	0.10	0.76	−0.32	−2.86	1.58	−0.51
	TN	0.45	0.03	0.83	0.11	−0.45	0.69	−0.28
	TP	0.47	0.03	0.85	1.06	−1.10	4.86	0.25
L7	DRP	0.31	0.06	0.58	0.10	−0.35	0.48	−0.26
	NO_3-N	0.52	0.10	0.80	0.10	−0.31	1.29	−0.34
	TN	0.47	0.03	0.83	0.08	−0.50	0.54	−0.21
	TP	0.53	0.02	0.88	0.06	−13.13	2.54	−0.18

TN, respectively, and exceeded 30% at 23 and 33 sites for NO_3-N and TP, respectively. For TP, the L7 method generally produced the largest loads (Figure 5). Loads estimated using L7 exceeded the loads estimated by the L5 and Ratio methods at 71 and 53 of the 77 sites, respectively, and by more than 30% at 48 and 32 sites, respectively. For NO_3-N, the L5 method generally produced the largest loads (Figure 5). Loads estimated using L5 exceeded the loads estimated by the L7 and Ratio methods at 58 and 62 of the 77 sites, respectively, and by more than 30% at 22 and 23 sites, respectively.

The proportion of significant differences in site loads calculated using different methods was least for TN and DRP (Figure 6). For DRP and TN, the proportion of significant differences in site loads were greatest for contrasts between the L7 and L5 methods, which were 51 and 56%, respectively (Figure 6). The remaining method contrasts produced significant differences for fewer than 31% of sites for DRP and TN. For TP, the L7 method produced significantly different loads to the L5 methods at 74% of the sites. The majority of the significant differences were larger L7 loads compared to the L5 loads (Figures 5 and 6). The remaining method contrasts produced significant differences for fewer than 36% of sites for TP. All method contrasts produced significantly different NO_3-N loads for at least 40% of the sites.

The flow-weighted residuals (ε) of the L5 rating curve had strong positive relationships with the differences between Ratio and L5 and L7 and L5 for all nutrient species (Table 3). This indicates that when the model underlying the L5 rating curve had a poor fit to the high-flow observations, the method tended to produce loads that were significantly different to L7 and the Ratio method. There were fewer and weaker relationships between the differences in the load estimates and the mean-weighted residuals of the L7 rating curve. This was consistent with the observation that ε values were generally lower for the L7 model (Table 2). The adjusted-r^2 values for both the L5 and L7 models were only weakly associated with the differences in loads calculated with the Ratio and L5 methods and the L5 and L7 methods for all nutrient species (Table 3).

Representativeness of Load Estimates

The short-term load estimates (*i.e.*, estimated using 5, 10, and 15 year subsamples) were variable and standard deviations of the 20 load estimates (*i.e.*, SE_R) varied considerably across the 77 sites (Figure 7). For example, site values of SE_R for TP for the 10-year duration and the L5 and L7 methods ranged from 0.01 to 1.8 $\log_{10}$ kg/ha/year (1-63 kg/ha/year) and 0.01 to 0.3 $\log_{10}$ kg/ha/year (1-2 kg/ha/year), respectively (Figure 7).

There were patterns in SE_R associated with durations, nutrient species, and calculation methods (Figure 7). Site values of SE_R were larger for shorter durations, and were generally least for the Ratio method and largest for the L5 method (Figure 7). The SE_R values were generally larger for TP compared to values for the other nutrient species calculated for the same duration and method.

For any sample period duration, the 95% confidence interval associated with representativeness was generally wider than the 95% confidence for precision (Figure 8, Table 4). For example, the characteristic precision (SE_P) for a DRP load calculated using the Ratio method for a 5-year period, and for the site with the median 95% confidence interval width, was plus or minus 21% (Table 4). By contrast, the 95% confidence intervals for the equivalent representativeness (SE_R) lay in the range −29%-41%. In addition, for DRP based on a 5-year sampling period and the Ratio method, 92% of sites had representativeness estimates that were larger than the characteristic precision (Table 4).

The proportion of sites having 95% confidence interval for representativeness that exceeded the

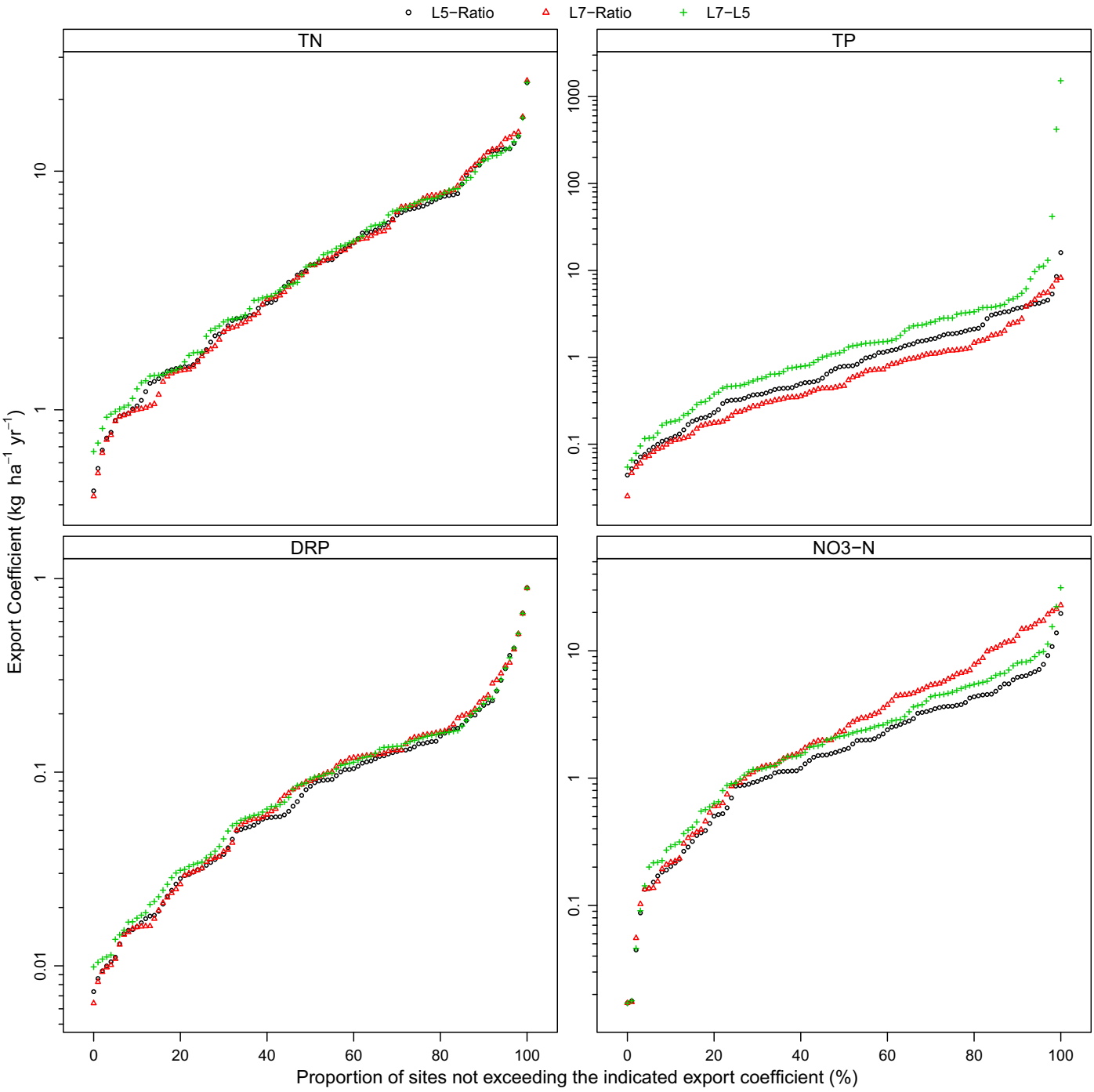

FIGURE 3. The Cumulative Frequency Distributions of the Estimated Site Loads, Expressed as Export Coefficients, for the 77 Sites Estimated from the 20-Year Dataset Using Three Load Calculation Methods. TN, Total Nitrogen; TP, Total Phosphorus; DRP, Dissolved Reactive Phosphorus; NO3–N, nitrate-nitrogen.

95% confidence interval for precision decreased with increasing sample period duration for all nutrient species for the Ratio method (Table 4). There was also a pattern of decreasing proportion of sites with representativeness exceeding precision with increasing sample period duration for the L5 and L7 methods, but the proportion of representativeness estimates that exceeded precision was higher than for the Ratio method for all sample period durations (Table 4). At some sites, 95% confidence interval for representativeness were much larger than the precision estimates, particularly for TP and for the L5 and L7 methods (Figure 8, Table 4).

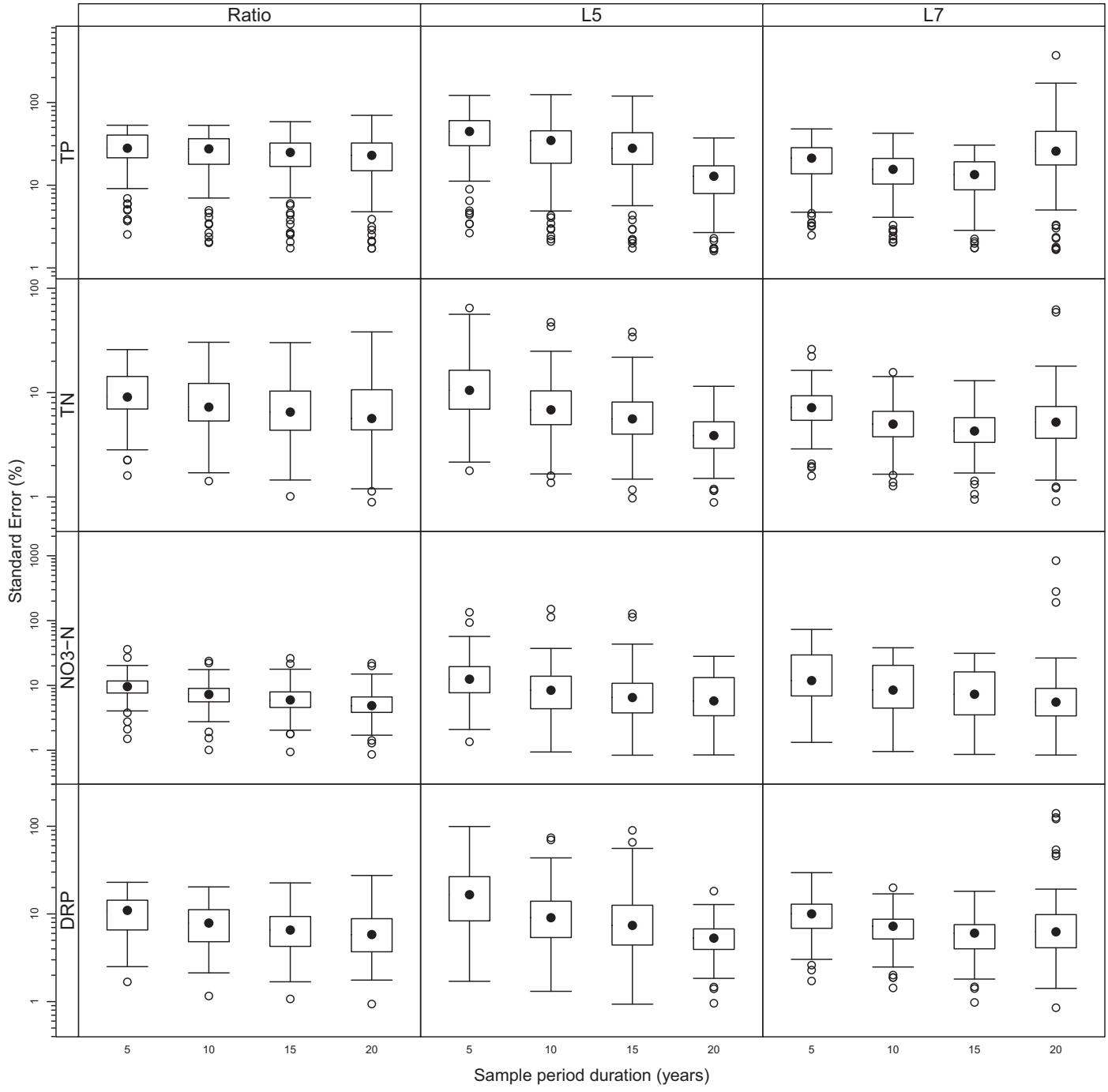

FIGURE 4. Distributions of the Standard Errors for the 77 Sites as Proportions of the Estimated Load (%) for Each Method and Nutrient Species. The box plots show the distributions of SE for the full 20-year dataset and the characteristic precision (SE_P; the mean of SE calculated for each of the 20 load estimates) for each of the three shorter sample period durations 5, 10, and 15 years. The box indicates the interquartile range, the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. Outliers are indicated by open circles.

Variation in Precision and Representativeness between Sites

Correlation between the five explanatory variables was generally low ($\tau < 0.4$). Exceptions to this included a correlation of 0.53 between γ_q and CV_q indicating that the skewness of the flow distribution was associated with inter annual flow variability and some correlations of around 0.5 between skewness of the concentration distribution (γ_c) and the absolute

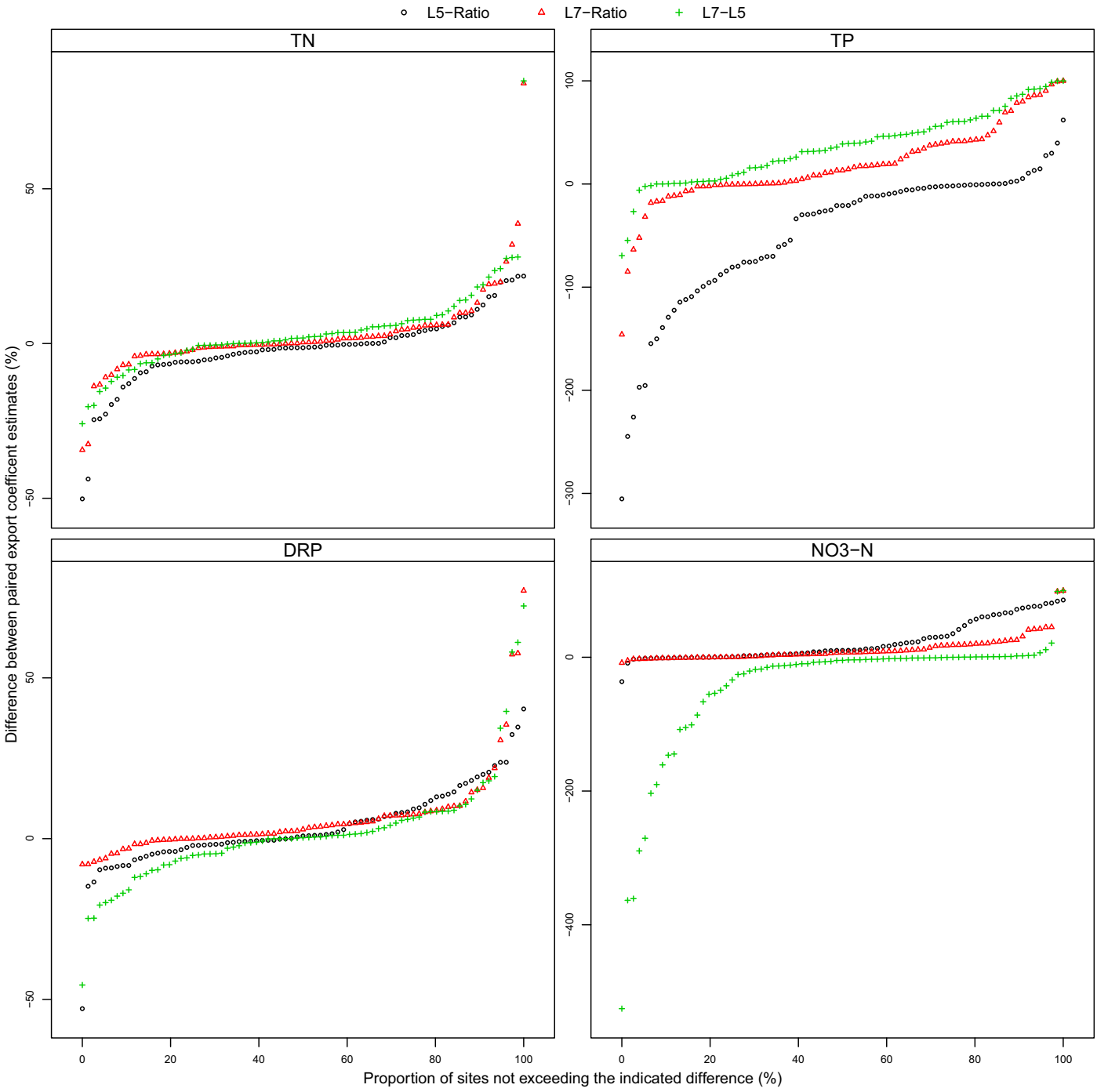

FIGURE 5. Cumulative Frequency Distributions of the Differences in Site Export Coefficients (%) at Each of the 77 Sites Calculated Using the Three Load Estimation Methods.

value of the mean flow-weighted residuals ($|\varepsilon|$). The strength of the concentration-flow relationship (adjusted-r^2) was weakly correlated with the other explanatory variables ($\tau < 0.2$). The skewness of the concentration and flow distributions (γ_q and γ_q) were only weakly correlated for all nutrients ($\tau < 0.1$).

The strength of the relationships between SE_P and SE_R and the explanatory variables varied by nutrient species and by method (Table 5). The relationships of both SE_P and SE_R with γ_c, γ_q, and CV_q were consistently significant and positive (*i.e.*, precision and representativeness decreased with increasing values of these explanatory variables).

Relationships of both SE_P and SE_R with adjusted-r^2 were always weak and were often not significant. When significant, relationships of SE_P and SE_R with

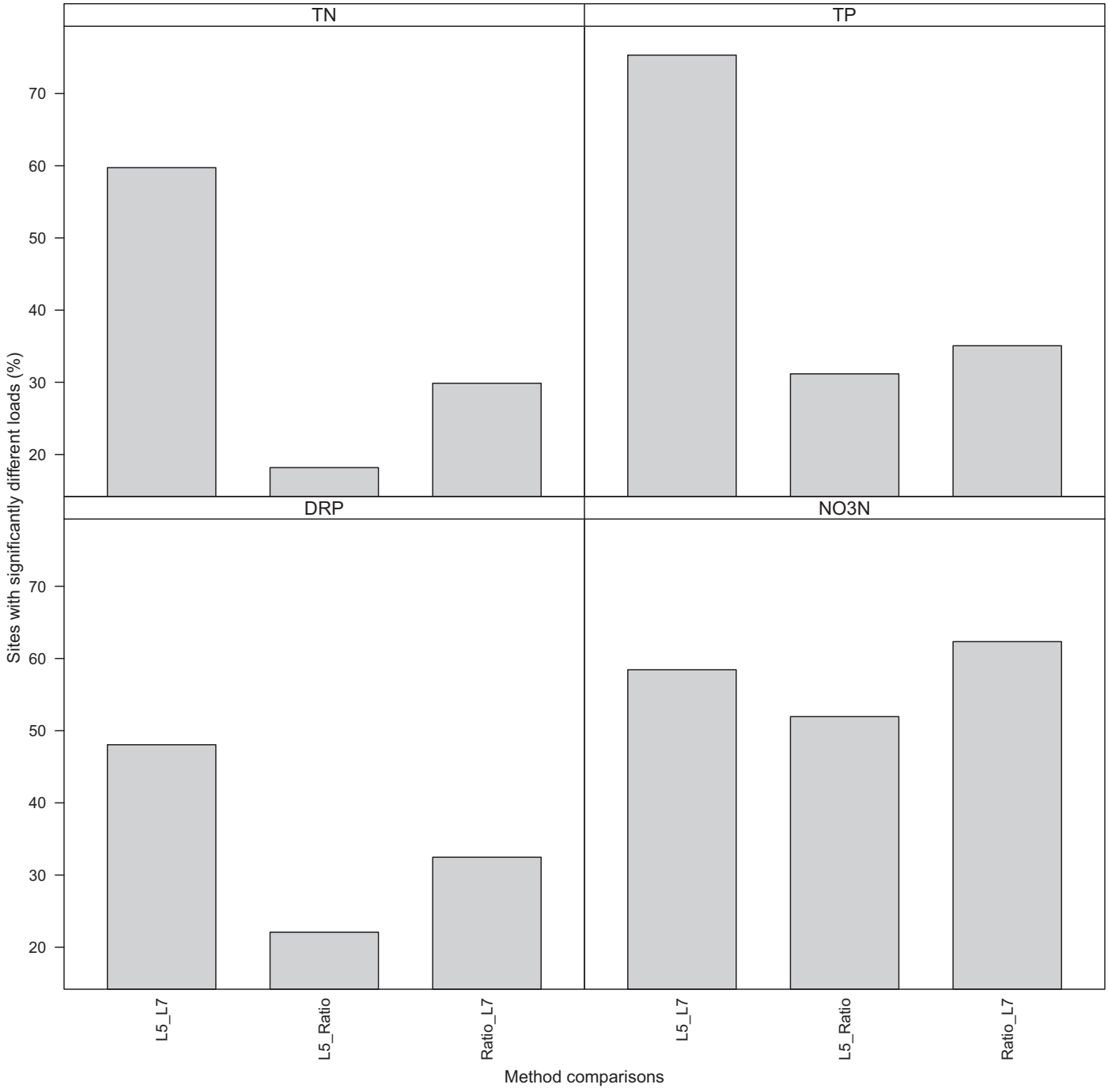

FIGURE 6. Proportion of Significant Differences for Loads Calculated Using Different Methods for the 77 Sites. The calculation methods that define the contrasts are shown in the *x*-axis labels.

r^2 were positive. This indicates that precision and representativeness decreased with increasing r^2. By contrast, relationships of SE_P and SE_R with the absolute value of the mean flow-weighted residuals ($|\varepsilon|$) were generally positive, indicating that for this measure, precision and representativeness increased with increasing model suitability.

DISCUSSION

This study examined differences in catchment nutrient loads calculated using three methods from monthly concentration data and daily flows at multiple sites. We were not able to assess the bias of our

TABLE 3. Kendall's Rank Correlation Coefficient (τ), Measuring the Association between the Differences in Load Magnitudes Calculated Using Different Methods and the Flow Weighted Residuals (ε) and the Adjusted-r^2 Values for the Regression Models That Define the Rating Curves for the L5 and L7 Models. The asterisks indicate τ statistics that were significant at the 5% level, na indicates that the explanatory variable (ε and r^2) were not applicable to the method contrast.

Nutrient Species	Method Contrast	Flow-Weighted Residuals (ε)		Adjusted-r^2	
		L5	L7	L5	L7
DRP	L7_L5	0.59*	0.12	−0.28*	−0.26*
	Ratio_L5	0.55*	na	−0.36*	na
	L7_Ratio	na	0.09	na	0.16*
NO_3-N	L7_L5	0.7*	0.15	−0.37*	−0.48*
	Ratio_L5	0.62*	na	−0.42*	na
	L7_Ratio	na	−0.05	na	0.30*
TN	L7_L5	0.63*	0.08	−0.14	−0.11
	Ratio_L5	0.62*	na	−0.29	na
	L7_Ratio	na	−0.31*	na	0.13
TP	L7_L5	0.59*	0.11	0.17*	0.30*
	Ratio_L5	0.71*	na	0.09	na
	L7_Ratio	na	0.00	na	0.20*

load estimates because we could not estimate the "true" load. However, the large number of sites and multiple nutrient species enabled us to examine several aspects of load calculation that have been rarely considered including: differences in loads estimated using alternate methods; the precision and representativeness of load estimates; and the underlying distributional and functional relationships that influence differences in load magnitudes associated with alternate methods, and their precision and representativeness.

Differences between Methods

In our study, differences in the magnitude of loads calculated using alternate methods were generally less than 30%, particularly for DRP and TN (Figure 5). For these species, differences between loads calculated using alternate methods were not statistically significant at many sites (Figure 6). The magnitude and proportion of significant differences in site loads estimated using alternate methods were greatest for NO_3-N and TP.

The reasons for differences in the magnitudes of loads estimated using different methods can be broadly understood in terms of the way the methods represent the functional relationship between flow and both concentration and unit loads and the characteristics of these relationships at the site. Relationships between flow and concentration can be curvilinear in log-log space and this was particularly the case for NO_3-N and TP (Figure 2). For NO_3-N, the rate of change in concentrations with flow could plateau or decrease at high flows at many sites. This is associated with initial increases in concentration in response to increased flow and subsequent dilution at high flows, which is consistent with the likely groundwater sources of NO_3-N (Woodward *et al.*, 2013). The quadratic term included in the L7 method is able to fit a rating curve that represents this curvilinear relationship, whereas the L5 method is unable to represent the reduction in concentration at high flows (Figure 2). This results in the tendency for L5 to predict higher site NO_3-N loads than the L7 or Ratio method (Figures 5 and 6). This tendency increased, the more the L5 rating curve failed to represent the observed values at high flows (*i.e.*, the larger the ε value; Table 3). We consider that at least some of the NO_3-N load estimates made using the L5 methods are unrealistically high as they exceeded the corresponding TN loads by large margins (Figure 3). In addition, five of the NO_3-N export coefficients calculated using the L5 method exceeded 20 kg/ha/year. These export coefficients are consistent with loss rates under intensive pastoral farming and are unlikely to be exceeded in large catchments with mixed land cover and where some attenuation of nitrogen is likely (McDowell and Wilcock, 2008).

By contrast to NO_3-N, there was a tendency for rate of change in TP concentrations to increase with flow (Figure 2). The inclusion of the quadratic term in the L7 method means that it is able to represent the curvilinear relationship between TP concentrations and flow in log-log space. The regression underlying the L5 rating curve cannot represent this curvilinear relationship (Figure 2). Because there were few observations at high flows, there was a tendency for the TP rating curves fitted with the L5 models to predict concentrations at high flows that were markedly lower than the observations (large ε values for TP at the majority of sites, Table 2). This meant that the L5 model produced load estimates that tend to be less than the L7 model (Figures 5 and 6) and this tendency increased the more the L5 rating curve failed to represent the observed values at high flows (*i.e.*, the larger the ε value, Table 3). It is therefore reasonable to assume that the L5 method frequently underestimated estimates of the TP loads.

The TP export coefficients produced by the L7 method tended to be significantly larger than those produced by the Ratio method. Contrary to the rating methods, we did not develop a method for quantifying the degree to which the Ratio model assumptions were violated (or the lack of fit of the Ratio model to the observations). However, inspection of the plots of unit loads *vs.* flow indicated that these relationships were frequently nonlinear (Figure 2), indicating that the first assumption of the Ratio method was

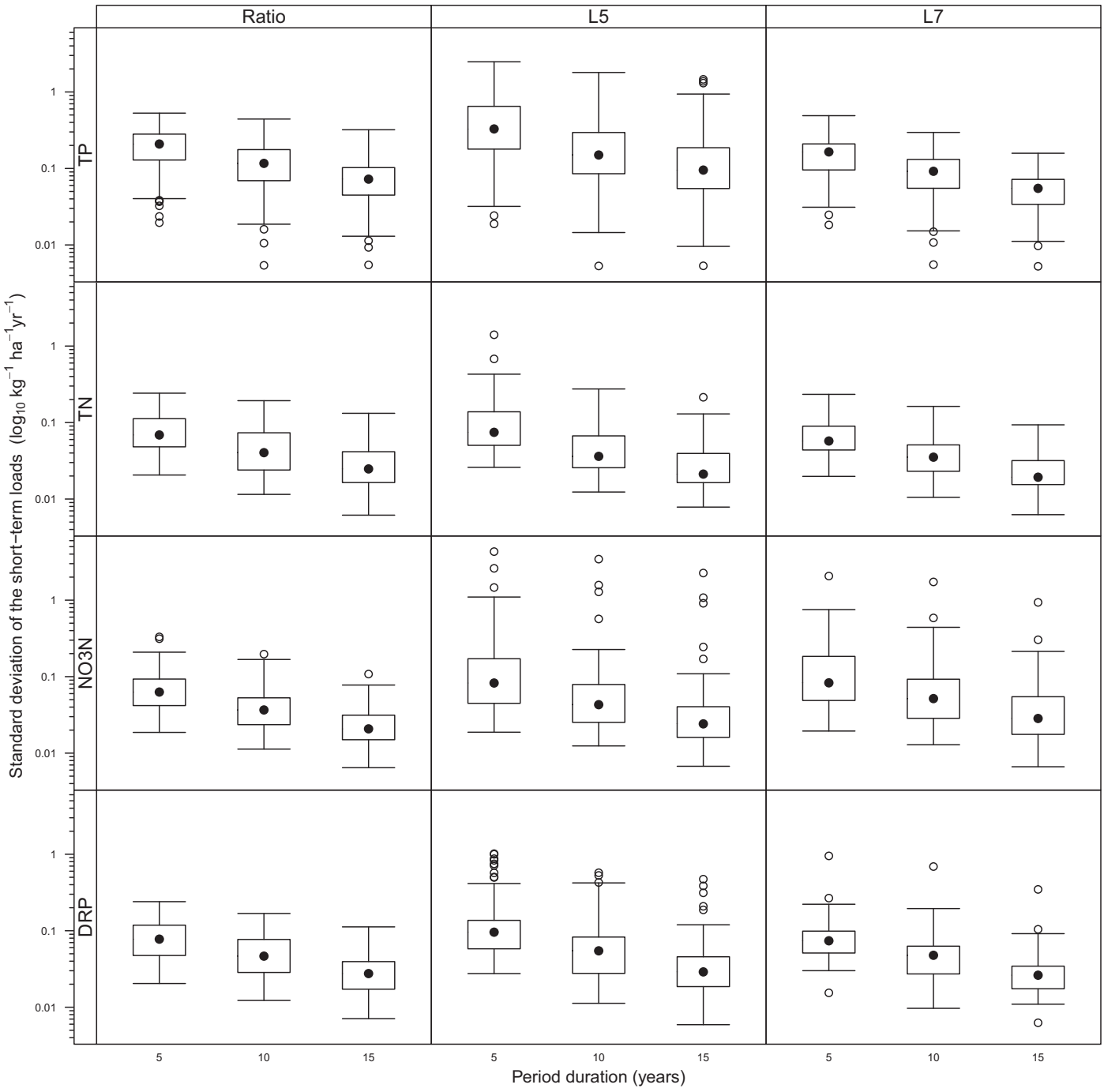

FIGURE 7. Distributions of the Standard Deviations of the Short-Term Load Estimates (SE_R, $\log_{10}$ kg/ha/year) for Each Method and Nutrient Species for the 77 Sites by Sample Period Duration (x-axis). The box indicates the inter-quartile range, the dot within the box indicates the median and 95% of the data lies within the whiskers. Outliers are indicated by open circles.

violated. It is therefore reasonable to assume that the TP loads estimated using the Ratio method was frequently biased.

These findings suggest that the L7 method is generally more appropriate, particularly when the relationships between flow and concentration in log-log space are curvilinear. However, some of the export coefficients estimated using the L7 method for TP were extremely high (*e.g.*, three were greater than 10 kg/ha/year; Figure 3) and probably unrealistic. The L7 model can introduce curvature that may result in very extreme estimates near or beyond the limits of the sampled values of flow or time in the dataset (Hirsch, 2014). This is particularly likely when the observations are sparse or absent at high flows, as was the case with our data.

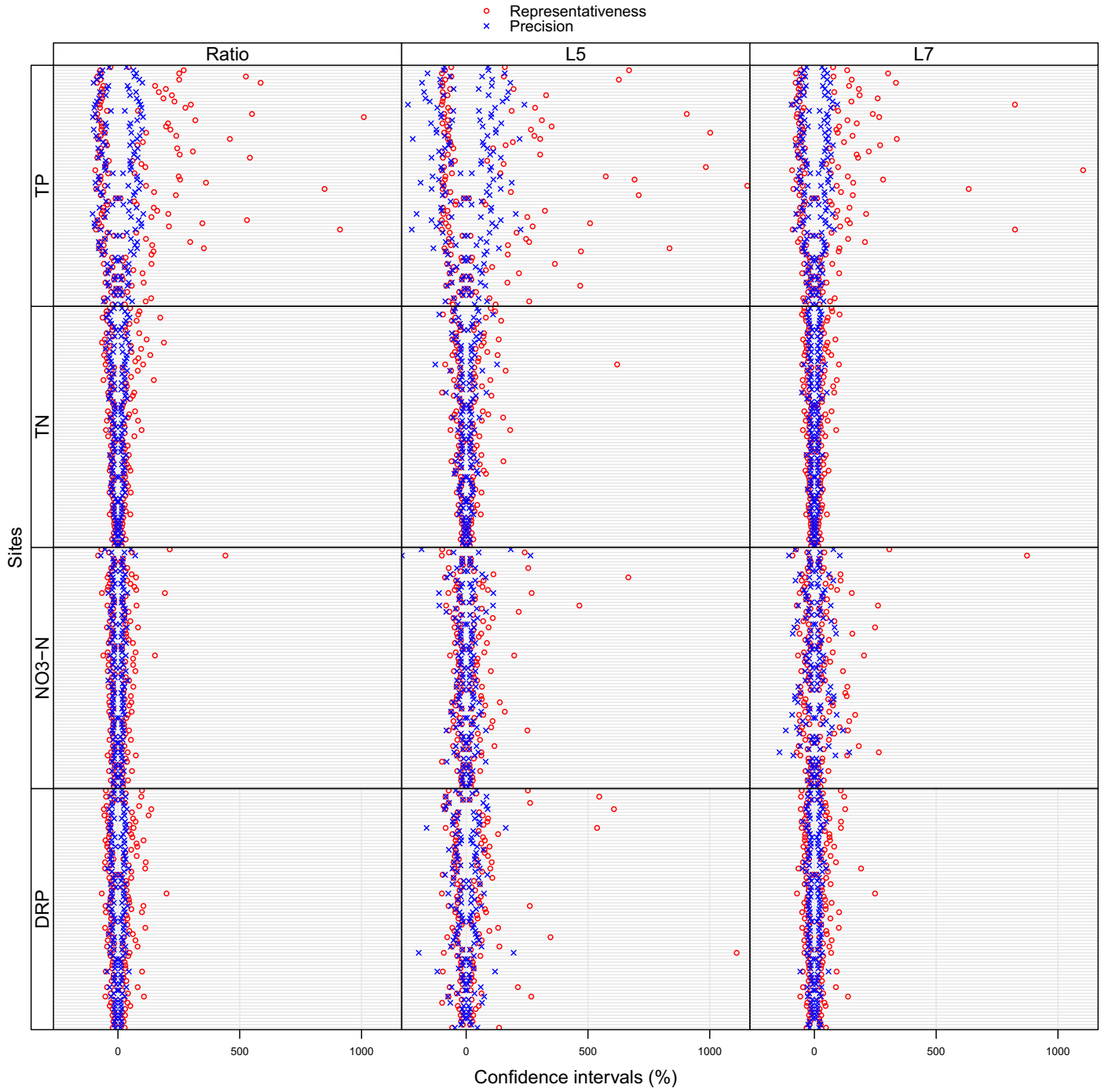

FIGURE 8. Relative 95% Confidence Intervals for Precision and Representativeness (Equation 8) of Load Estimates for Each Method and Nutrient Species for the 5-Year Time Period Nutrient. The upper and lower confidence intervals for characteristic precision (CI_P) and representativeness (CI_R) are indicated as pairs of points for each of the 77 sites (y-axis). The sites have been ordered by their respective γ_c values. The x-axis has been truncated and some upper confidence intervals exceeded 1000%.

Precision

Imprecision of load estimates occurs because monthly river concentration data are subject to considerable sampling error. Sampling error increases with reducing sample size; therefore reducing sample period duration. Imprecision also increases when the distribution of the concentration data is strongly skewed (Table 5) because the tail of the concentration distribution is poorly characterized. If there is a strong positive relationship between concentration and flow, the poorly characterized individual high-flow samples strongly influence the load estimate (Table 5).

TABLE 4. Summary of the Characteristic Precision (P) and Representativeness (R) Measures for the Site Export Coefficient Estimates Calculated for Different Methods, Sampling Period Durations, and Nutrient Species. The P and R values are the 95% confidence interval for precision (calculated from SE_P) and representativeness (calculated from SE_R: Equation 8) having median width (over the 77 sites). The third column for each method shows the proportion of sites for which the width of the 95% confidence interval for representativeness exceeded the 95% confidence interval for precision.

Nutrient Species	Sampling Period Duration	Ratio			L5			L7		
		P (%)	R (%)	R > P (%)	P (%)	R (%)	R > P (%)	P (%)	R (%)	R > P (%)
DRP	5	−21; 21	−29; 41	92	−38; 38	−36; 55	81	−21; 21	−30; 43	94
	10	−16; 16	−19; 23	73	−22; 22	−23; 29	62	−15; 15	−18; 22	79
	15	−13; 13	−11; 12	36	−15; 15	−13; 15	36	−12; 12	−12; 13	43
NO_3-N	5	−20; 20	−24; 32	83	−31; 31	−33; 50	79	−23; 23	−31; 44	69
	10	−15; 15	−15; 18	58	−18; 18	−17; 21	64	−18; 18	−19; 23	61
	15	−12; 12	−9; 10	29	−13; 13	−11; 12	39	−15; 15	−11; 13	35
TN	5	−19; 19	−24; 32	90	−23; 23	−29; 40	79	−15; 15	−22; 29	92
	10	−15; 15	−14; 17	64	−16; 16	−15; 17	61	−11; 11	−13; 15	71
	15	−13; 13	−9; 10	23	−12; 12	−9; 10	27	−9; 9	−7; 8	43
TP	5	−58; 58	−59; 147	97	−126; 126	−74; 287	83	−43; 43	−50; 98	94
	10	−56; 56	−41; 69	57	−80; 80	−49; 98	61	−35; 35	−32; 47	71
	15	−50; 50	−28; 40	18	−60; 60	−34; 52	38	−28; 28	−21; 26	39

TABLE 5. Kendall's Rank Correlation Coefficient (τ), Measuring the Association of the SE_P (first value) and SE_R (in parentheses) with Five Explanatory Variables for the Four Nutrient Species and Three Load Calculation Method Combinations for the 10-Year Sample Period Duration. Results were very similar for the 5 and 15-year sample period durations. The asterisks indicate τ statistics that were significant at the 5% level, na indicates that the explanatory variable was not applicable to the method.

Nutrient Species	Method	Adjusted-r^2	γ_c	γ_q	CV_q	$\lvert\varepsilon\rvert$
DRP	Ratio	na	0.51* (0.36*)	0.22* (0.2*)	0.3* (0.33*)	na
	L5	−0.07 (0.01)	0.4* (0.32*)	0.25* (0.25*)	0.29* (0.3*)	0.35* (0.31*)
	L7	0.04 (0.04)	0.42* (0.31*)	0.23* (0.22*)	0.35* (0.32*)	0.46* (0.32*)
NO_3-N	Ratio	na	0.54* (0.43*)	0.23* (0.23*)	0.28* (0.28*)	na
	L5	0.12 (0.1)	0.34* (0.34*)	0.3* (0.24*)	0.31* (0.27*)	−0.08 (0.01)
	L7	0.27* (0.17*)	0.22* (0.28*)	0.33* (0.33*)	0.32* (0.42*)	0.25* (0.23*)
TN	Ratio	na	0.5* (0.53*)	0.22* (0.21*)	0.17* (0.2*)	na
	L5	0.06 (0.05)	0.45* (0.44*)	0.32* (0.33*)	0.25* (0.33*)	0.43* (0.37*)
	L7	0.09 (0.1)	0.46* (0.38*)	0.26* (0.31*)	0.21* (0.34*)	0.55* (0.29*)
TP	Ratio	na	0.53* (0.5*)	0.21* (0.17*)	0.08 (0.09)	na
	L5	0.25* (0.2*)	0.41* (0.37*)	0.32* (0.31*)	0.15* (0.16*)	0.49* (0.42*)
	L7	0.33* (0.34*)	0.45* (0.43*)	0.21* (0.2*)	0.16* (0.16*)	0.37* (0.23*)

The rating methods tend to have lower precision than the Ratio method because individual observations can exert large leverage on the definition of the rating curve. In particular, infrequent high concentration observations that are associated with high flows will strongly influence the definition of the high-flow portion of rating. This means that individual observations have a large influence on the magnitude of individual unit loads, leading to large variance (low precision) of the load estimates (Vieux and Moreda, 2003).

The Ratio method is also affected by poor characterization of the tail of the concentration distribution. Because the method is based on the mean of the unit loads, the weight that any individual observation has on the load estimate is reduced compared to the Rating method (Young *et al.*, 1988). However, in our study, loads estimated using the Ratio method at some sites had lower precision (larger SE_P values) than those estimated using the L5 and L7 methods, particularly for TP (Figure 4). We consider this is likely to arise if a small number of high-flow observations are associated with high concentrations. In these circumstances, individual samples can have a large influence on the magnitude of the mean unit load leading to high variance of the load estimates. These observations induce nonlinearity in the unit load-flow relationship (Figure 2) and therefore violate the underlying assumptions of the Ratio method.

Representativeness of Load Estimates Made from Short-Term Datasets

Variation in the short-term mean annual catchment load estimates (SE_R) was considerable,

indicating that estimates made for specific time periods poorly represent the characteristic mean annual catchment load. For a given site and sampling period, SE_R was generally larger than the characteristic precision (SE_P; Figure 8, Table 4). Thus, load estimates are likely to fluctuate to a greater extent than indicated by their associated precision estimates. This means that precision is likely to be an "optimistic" assessment of the actual repeatability of load estimates and detectable differences in loads are likely to be larger than indicated by their precision.

The high variability associated with loads calculated for short time periods is probably associated with two main factors: the representativeness of the concentration data and temporal variation in environmental conditions. Concentration samples associated with short time periods poorly represent the concentration distribution. Between-site variation in representativeness (SE_R) of load estimates was strongly positively associated with the skewness of the concentration distribution (γ_c) and flow distribution (γ_q). This is because, for any fixed number of random concentration observations, the sampling error will be lower, and the functional relationships between concentration and flow more accurately characterized, when distributions are more normal. Loads calculated for sites with strongly skewed distributions are sensitive to individual high-flow observations, particularly when the concentration-flow relationship is strong and this leads to high variance in the estimates. This is particularly the case for TP and reflects its dominant transport mechanism being associated with episodic runoff and high flows (McDowell *et al.*, 2014).

We note that some sites had very large SE_R values for the L5 and L7 methods, particularly for TP. These large values are associated with the leverage of individual high-flow observations and the potential for rating methods to produce extreme load estimates when ratings are extended beyond the limits of the sampled data (Hirsch, 2014). It is likely that the rating curves underlying these individual load estimates would not be considered credible, were they to be inspected as part of a load calculation procedure.

Contaminant fluxes are also influenced by temporal variation in hydrology and the processes that mobilize contaminants in the catchment. Hydrological studies have shown considerable interannual variation in hydrological regimes in New Zealand rivers (Booker, 2013; McKerchar and Henderson, 2003; Mullan, 1995). This variation is subject to temporal correlation that is partly explained by the El Niño–Southern Oscillation (ENSO), which has a characteristic period length of five years (Mullan, 1995). This means that particular sampling periods of several years may be associated with flows that vary significantly to another sampling period leading to differences in mean flows and flow distributions with consequent effects on load estimates. Scarsbrook *et al.* (2003) also suggested that ENSO has effects on contaminant concentrations that are not a direct consequence of flow variation because trends in water quality were consistent with trends in ENSO even when concentration data were adjusted for flow.

To test the effect of temporal variation in flows, we also estimated the SE_R values for each time period using the flow data for the entire flow record. This simulates situations in which a short period of concentration data can be associated with flow data that represents the long term; either longer time series or flow estimates based on regionalization or other models (*e.g.*, Snelder and Booker, 2013; Woods *et al.*, 2006). Using the entire 20 years of flow data to estimate loads resulted in small reductions in SE_R for all time periods. However, the 95% confidence intervals for representativeness exceeded precision for the majority of sites (*i.e.*, >50%) for most time periods and methods. This indicates that it is sampling error, rather than temporal variation in environmental conditions, that is the largest contributor to the low representativeness of load estimates.

Limitations

We acknowledge some limitations of our study. In particular, our results pertain to catchments of a size similar to those in the NRWQN (Table 1). Different factors and sources of uncertainty are likely to influence load estimates made for catchments with different characteristics. For instance, concentration-flow relationships are likely to vary with catchment area due to differences in attenuation and removal processes, even when land use is taken into account (Alexander *et al.*, 2007).

Use of daily mean flow can bias load estimates by ignoring flow variation within days (Robertson and Roerish, 1999). We fitted the L5 and L7 rating curves to data pertaining to both instantaneous and daily flows associated with the concentration samples. Rating curves were reasonably insensitive to whether daily or instantaneous flows were used. Differences in adjusted-r^2 values across sites and nutrient species were also small (mean <1%, standard deviation <6%). These results are consistent with the majority of sites being located on large rivers for which intra-day flow variation is small. We therefore consider that our findings are robust, particularly given that they are derived from comparisons made under fixed assumptions and data.

Recommendations for Regulatory Authorities

Some authors have suggested that rating methods should be used where the concentration-flow relationship is strong (*e.g.*, $r^2 > 50\%$; Quilbé *et al.*, 2006). Our results indicate that the strength of the concentration-flow relationship is an insufficient criterion to judge the appropriateness of the rating methods when using monthly data. We found that rating curves could have high r^2 values; but not adequately represent aspects of the concentration-flow relationship that were evident in the data. In particular, we showed that large differences in loads calculated using different methods and low precision and representativeness were weakly related to r^2 but were strongly related to the flow-weighted residuals (ε) (Table 3). In addition, for the rating methods, r^2 was positively related to SE_P and SE_R (*i.e.*, precision and representativeness decreased with increasing r^2, Table 5), which at face value is inconsistent with the advice to use rating methods when the relationship between flow and concentration is strong (*e.g.*, Defew *et al.*, 2013; Preston *et al.*, 1989; Young *et al.*, 1988). This occurred in our study because our monthly data poorly represented the high flows. Apparently strong relationships between concentration and flow arose due to a "pan handle" effect where a small number of high-flow samples have large leverage on the concentration-flow relationship. Load estimates made from these data have high variance due to their sensitivity to the individual high-flow samples and this results in low precision and representativeness.

Our results indicate that the L7 method or a similar (*i.e.*, nonlinear) regression method is generally preferable for constructing rating curves as these are able to accommodate the frequently curvilinear relationships between concentration and flow (in log space). However, in some circumstances, a flexible model will produce poor load estimates and there is no single preferred model. Weighted Regression on Time Discharge and Season (WRTDS) (Hirsch, 2014) provides a flexible method for fitting rating curves that can be more robust than methods tested in this study. In particular, WRTDS is more robust to heteroskedastic regression residuals that can lead to biased estimates using the L5 and L7 models (Hirsch, 2014). However, the WRTDS method requires at least 120 concentration observations and would be prone to bias when the concentration-flow distribution is poorly characterized (Hirsch, 2014).

We recommend that the adequacy of any method needs to be judged by considering whether the model underlying the load calculation method is consistent with the available data and the expected behavior of the contaminant. This highlights the importance of inspecting concentration-flow, unit load-flow plots, and regression residuals. Our results also highlight the importance of thorough examination of the data when calculating loads and the inadvisability of "unsupervised" load estimation either by automation or inexperienced analysts.

When using monthly data to estimate loads, high-flow samples will be poorly represented and this can lead to low precision and representativeness (Vieux and Moreda, 2003). We suggest that this situation can only be improved by supplementing monthly observations with more data, particularly high-flow observations. Our results indicate that the precision and representativeness of all methods decrease with increasing skewness of the concentration and flow distributions, increasing variation in annual mean flows and the mean flow weighted residual value (Table 5). These results provide a means to identify the sites and nutrient species that are likely to have the least precise and representative load estimates and can be used to prioritize sites for collection of additional concentration samples.

CONCLUSION

Nutrient load estimates made from short sample period durations of monthly concentration data tend to be more variable than indicated by their estimated precision, irrespective of the load calculation method. Thus, detectable differences in loads calculated from monthly data are generally larger than indicated by precision estimates. In our study, differences in loads estimated using alternative methods were frequently not statistically significant despite relatively long sample records (*i.e.*, 20 years). However, significant differences in loads estimated using different methods can often be understood as a consequence of poor agreement between the data and the model underlying one of the methods. Choice of load calculation method should therefore be based on careful consideration of the distributional and functional characteristics of concentration and flow. Differences in loads estimated using alternative methods and the precision and representativeness of any estimate are disproportionately affected by high-flow observations, which are often limited when concentration data is based on infrequent sampling. Thus, the choice method is not as important as considering whether the data are sufficient to estimate loads that are fit for purpose (*i.e.*, of sufficient precision and accuracy).

ACKNOWLEDGMENTS

This work was partly funded by the New Zealand Ministry for Business, Innovation and Employment's Clean Water, Productive Land programme (contract C10X1006) and Wheel of Water Programme (contract ALNC1102). Access to concentration and flow data was made available by the National Institute of Water and Atmospheric Research. We thank Greg Barkle and Graham McBride and three anonymous reviewers whose comments improved the original manuscript.

LITERATURE CITED

Alexander, R.B., E.W. Boyer, R.A. Smith, G.E. Schwarz, and R.B. Moore, 2007. The Role of Headwater Streams in Downstream Water Quality. Journal of the American Water Resources Association 43:41-59.

Alexander, R.B., A.H. Elliott, U. Shankar, and G.B. McBride, 2002. Estimating the Sources and Transport of Nutrients in the Waikato River Basin, New Zealand. Water Resources Research 38:1268.

Beale, E.M.L., 1962. Some Uses of Computers in Operational Research. Industrielle Organisation 31:27-28.

Booker, D.J., 2013. Spatial and Temporal Patterns in the Frequency of Events Exceeding Three Times the Median Flow (FRE3) across New Zealand. Journal of Hydrology (New Zealand) 52:15.

Cochran, W.G., 1977. Sampling Techniques. John Wiley & Sons, New York.

Cohn, T.A., 2005. Estimating Contaminant Loads in Rivers: An Application of Adjusted Maximum Likelihood to Type 1 Censored Data. Water Resources Research 41. http://onlinelibrary.wiley.com/doi/10.1029/2004WR003833/full, *accessed* January 2016.

Cohn, T.A., D.L. Caulder, E.J. Gilroy, L.D. Zynjuk, and R.M. Summers, 1992. The Validity of a Simple Statistical Model for Estimating Fluvial Constituent Loads: An Empirical Study Involving Nutrient Loads Entering Chesapeake Bay. Water Resources Research 28:2353-2363.

Cohn, T.A., L.L. Delong, E.J. Gilroy, R.M. Hirsch, and D.K. Wells, 1989. Estimating Constituent Loads. Water Resources Research 25:937-942.

Daily, G.C., S. Alexander, P.R. Ehrlich, L. Goulder, J. Lubchenco, P.A. Matson, H.A. Mooney, S. Postel, S.H. Schneider, D. Tilman, and G.M. Woodwell, 1997. Ecosystem Services: Benefits Supplied to Human Societies by Natural Ecosystems. Issues in Ecology 2:2-15.

Davies-Colley, R.J., D.G. Smith, R.C. Ward, G.G. Bryers, G.B. McBride, J.M. Quinn, and M.R. Scarsbrook, 2011. Twenty Years of New Zealand's National Rivers Water Quality Network: Benefits of Careful Design and Consistent Operation. Journal of the American Water Resources Association 47:750-771.

Defew, L.H., L. May, and K.V. Heal, 2013. Uncertainties in Estimated Phosphorus Loads as a Function of Different Sampling Frequencies and Common Calculation Methods. Marine and Freshwater Research 64:373-386.

Dolan, D.M., A.K. Yui, and R.D. Geist, 1981. Evaluation of River Load Estimation Methods for Total Phosphorus. Journal of Great Lakes Research 7:207-214.

Duan, N., 1983. Smearing Estimate: A Nonparametric Retransformation Method. Journal of the American Statistical Association 78:605-610.

Efron, B., 1981. Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods. Biometrika 68:589-599.

Environment Agency, 2013. Livestock Slurry Storage in Nitrate Vulnerable Zones. Environment Agency, Rotherham, UK.

Hirsch, R.M., 2014. Large Biases in Regression-Based Constituent Flux Estimates: Causes and Diagnostic Tools. Journal of the American Water Resources Association 50:1401-1424.

Iowa State Legislature, 2010. Chapter 65: Animal Feeding Operations. Iowa State Legislature, Des Moines, Iowa.

Johnes, P.J., 2007. Uncertainties in Annual Riverine Phosphorus Load Estimation: Impact of Load Estimation Methodology, Sampling Frequency, Baseflow Index and Catchment Population Density. Journal of Hydrology 332:241-258.

Kronvang, B. and A.J. Bruhn, 1996. Choice of Sampling Strategy and Estimation Method for Calculating Nitrogen and Phosphorus Transport in Small Lowland Streams. Hydrological Processes 10:1483-1501.

Larned, S.T., M.R. Scarsbrook, T. Snelder, N.J. Norton, and B.J.F. Biggs, 2004. Water Quality in Low-Elevation Streams and Rivers of New Zealand. New Zealand Journal of Marine & Freshwater Research 38:347-366.

Littlewood, I.G., C.D. Watts, and J.M. Custance, 1998. Systematic Application of United Kingdom River Flow and Quality Databases for Estimating Annual River Mass Loads (1975-1994). Science of the Total Environment 210:21-40.

McDowell, R.W., R.M. Dils, A.L. Collins, K.A. Flahive, A.N. Sharpley, and J. Quinn, 2015. A Review of the Policies and Implementation of Practices to Decrease Water Quality Impairment by Phosphorus in New Zealand, the UK, and the US. Nutrient Cycling in Agroecosystems 104:289-305.

McDowell, R.W., P. Moreau, J. Salmon-Monviola, P. Durand, P. Leterme, and P. Merot, 2014. Contrasting the Spatial Management of Nitrogen and Phosphorus for Improved Water Quality: Modelling Studies in New Zealand and France. European Journal of Agronomy 57:52-61.

McDowell, R.W. and R.J. Wilcock, 2008. Water Quality and the Effects of Different Pastoral Animals. New Zealand Veterinary Journal 56:289-296.

McKerchar, A.I. and R.D. Henderson, 2003. Shifts in Flood and Low-Flow Regimes in New Zealand Due to Interdecadal Climate Variations. Hydrological Sciences Journal-Des Sciences Hydrologiques 48:637-654.

Meals, D.W., 1996. Watershed-Scale Response to Agricultural Diffuse Pollution Control Programs in Vermont, USA. Water Science and Technology 33:197-204.

Ministry for the Environment, 2014. National Policy Statement for Freshwater Management. Wellington, New Zealand.

Mullan, A.B., 1995. On the Linearity and Stability of Southern Oscillation-Climate Relationships for New Zealand. International Journal of Climatology 15:1365-1386.

Parliamentary Commissioner for the Environment, 2013. Water Quality in New Zealand: Land Use and Nutrient Pollution. Parliamentary Commissioner for the Environment, Wellington, New Zealand. http://www.pce.parliament.nz/assets/Uploads/PCE-Water-quality-land-use-web-ammended.pdf, *accessed* November 2016.

Phillips, J.M., B.W. Webb, D.E. Walling, and G.J.L. Leeks, 1999. Estimating the Suspended Sediment Loads of Rivers in the LOIS Study Area Using Infrequent Samples. Hydrological Processes 13:1035-1050.

Preston, S.D., V.J. Bierman, and S.E. Silliman, 1989. An Evaluation of Methods for the Estimation of Tributary Mass Loads. Water Resources Research 25:1379-1389.

Quilbé, R., A.N. Rousseau, M. Duchemin, A. Poulin, G. Gangbazo, and J.-P. Villeneuve, 2006. Selecting a Calculation Method to Estimate Sediment and Nutrient Loads in Streams: Application to the Beaurivage River (Québec, Canada). Journal of Hydrology 326:295-310.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/, *accessed* November 2016.

Richards, R.P. and J. Holloway, 1987. Monte Carlo Studies of Sampling Strategies for Estimating Tributary Loads. Water Resources Research 23:1939-1948.

Robertson, D.M. and E.D. Roerish, 1999. Influence of Various Water Quality Sampling Strategies on Load Estimates for Small Streams. Water Resources Research 35:3747-3759.

Roygard, J.K.F., K.J. McArthur, and M.E. Clark, 2012. Diffuse Contributions Dominate over Point Sources of Soluble Nutrients in Two Sub-Catchments of the Manawatu River, New Zealand. New Zealand Journal of Marine and Freshwater Research 46:219-241.

Scarsbrook, M.R., C.G. McBride, G.B. McBride, and G.G. Bryers, 2003. Effects of Climate Variability on Rivers: Consequences for Long Term Water Quality Analysis. Journal of the American Water Resources Association 39:1435-1447.

Smith, D.G. and R. Maasdam, 1994. New Zealand's National Water Quality Network. 1. Design and Physio-Chemical Characterisation. New Zealand Journal of Marine & Freshwater Research 28:19-35.

Smith, D.G. and G.B. McBride, 1990. New Zealand's National Water Quality Monitoring Network - Design and First Year's Operation. Water Resources Bulletin 26:767-775.

Snelder, T.H. and D.J. Booker, 2013. Natural Flow Regime Classifications Are Sensitive to Definition Procedures. River Research and Applications 29:822-838.

Vieux, B.E. and F.G. Moreda, 2003. Nutrient Loading Assessment in the Illinois River Using a Synthetic Approach. Journal of the American Water Resources Association 39:757-769.

Woods, R.A., J. Hendrikx, R. Henderson, and A. Tait, 2006. Estimating Mean Flow of New Zealand Rivers. Journal of Hydrology (New Zealand) 45:95-110.

Woodward, S.J., R. Stenger, and V.J. Bidwell, 2013. Dynamic Analysis of Stream Flow and Water Chemistry to Infer Subsurface Water and Nitrate Fluxes in a Lowland Dairying Catchment. Journal of Hydrology 505:299-311.

Young, T.C., J.V. DePinto, and T.M. Heidtke, 1988. Factors Affecting the Efficiency of Some Estimators of Fluvial Total Phosphorus Load. Water Resources Research 24:1535-1540.

Zar, J.H., 1999. Biostatistical Analysis. Prentice-Hall, Upper Saddle River, New Jersey.