

A MONTE CARLO TEST OF LOAD CALCULATION METHODS, LAKE TAHOE BASIN, CALIFORNIA-NEVADA¹

Robert Coats, Fengjing Liu, and Charles R. Goldman²

ABSTRACT: The sampling of streams and estimation of total loads of nitrogen, phosphorus, and suspended sediment play an important role in efforts to control the eutrophication of Lake Tahoe. We used a Monte Carlo procedure to test the precision and bias of four methods of calculating total constituent loads for nitrate-nitrogen, soluble reactive phosphorus, particulate phosphorus, total phosphorus, and suspended sediment in one major tributary of the lake. The methods tested were two forms of the Beale's Ratio Estimator, the Period Weighted Sample, and the Rating Curve. Intensive sampling in 1985 (a dry year) and 1986 (a wet year) provided a basis for estimating loads by the "worked record" method for comparison with estimates based on resampling actual data at the lower intensity that characterizes the present monitoring program. The results show that: (1) the Period Weighted Sample method was superior to the other methods for all constituents for 1985; and (2) for total phosphorus, particulate phosphorus, and suspended sediment, the Rating Curve gave the best results in 1986. Modification of the present sampling program and load calculation methods may be necessary to improve the precision and reduce the bias of estimates of total phosphorus loads in basin streams.

(KEY TERMS: aquatic ecosystems; statistical analysis; water quality; watershed management; Lake Tahoe; eutrophication; load calculation).

INTRODUCTION

Since 1980, the Lake Tahoe Interagency Monitoring Program (LTIMP) has been measuring stream discharge and concentrations of nutrients and sediment in up to ten tributary streams in the Lake Tahoe Basin, California-Nevada. The objectives of the LTIMP are "to acquire and disseminate the water-quality information necessary to support science-based environmental planning and decision making in the basin" (Boughton *et al.*, 1997). The LTIMP is a

cooperative program with support from 12 federal and state agencies with interests in the Tahoe Basin. The LTIMP data set is comprised of greater than 15,000 samples representing about 250 station-years of record for up to six water quality constituents. Its potential for research and decision-making has barely been tapped.

Concern about water quality in the Tahoe Basin is driven by the progressive eutrophication of the lake, which has been studied intensively since the early 1960s (Goldman, 2000). In spite of increased land-use controls and export of treated sewage effluent from the basin, primary productivity of the lake is increasing by more than 5 percent annually, and its clarity (measured by Secchi disk) is decreasing at an average rate of 0.25 m/yr. Until the early 1980s, nutrient limitation studies showed that primary productivity in the lake was nitrogen-limited. Now, after a half-century of accelerated nitrogen input (much of it from direct atmospheric deposition), the lake is phosphorus-limited (Chang *et al.*, 1992; Goldman *et al.*, 1993; Jassby *et al.*, 1995). Because the volume of the lake is so large (156 km³) and its hydraulic residence time so long (about 700 y; Jassby *et al.*, 1995), its eutrophication may be virtually irreversible. Since land use policies and water quality control programs in the basin are aimed largely at controlling or reducing the loads of nitrogen and phosphorus to the lake, it is important that tributary nutrient loads be accurately estimated (Reuter *et al.*, 1999).

Conceptually, the calculation of tributary mass loads requires evaluating an integral. The load in a given time interval between t_a and t_b is given by

¹Paper No. 01138 of the *Journal of the American Water Resources Association*. Discussions are open until February 1, 2003.

²Respectively, Principal, Hydroikos Associates, 2175 East Francisco Blvd., Suite A, San Rafael, California 94901; Graduate Student, INSTAAR and Department of Geography, University of Colorado, Boulder, Colorado 80309; and Professor, Department of Environmental Science and Policy, University of California, Davis, California 95616 (E-Mail/Coats: coats@hydroikos.com).

$$L = \int_{t_a}^{t_b} K \cdot Q_t \cdot C_t dt \quad (1)$$

where L is the total load in the time interval t_a to t_b ; K is a unit conversion factor; Q_t is the instantaneous discharge at time t ; and C_t is the instantaneous concentration at time t .

The instantaneous discharge can be measured by standard stream gaging techniques at the time of sampling, and continuous (or at least mean daily) discharge data are often readily available. The problem is that concentration of most constituents cannot be measured continuously, but has to be sampled and determined by chemical methods.

A number of methods and various refinements have been developed for estimating loads. These fall into three categories: averaging estimators, ratio estimators, and regression estimators (Preston *et al.*, 1989). Averaging estimators use averages of concentration and discharge for different time intervals, and sum the results over the water year. This approach may give biased results, if the sampling does not adequately characterize the extremes of flow and concentration during the entire year (Dolan *et al.*, 1981; Ferguson, 1987). The period-weighted sample method (PWS) (Dann *et al.*, 1986) is a type of averaging estimator. In this method, each two successive concentrations are averaged, multiplied by the cumulative discharge between sampling times, and the resulting load increments summed over the water year. The PWS has been used at Hubbard Brook to calculate total ion loads leaving the watersheds (Likens *et al.*, 1977).

The method of the worked record may also be thought of as an averaging method. In this method, the time trace of discharge and concentration are plotted together, and the mean daily concentration is interpolated for days on which samples were not collected. This allows the technician to adjust concentrations up or down to take account of discharge variation. With a good database and relatively low intradaily variability in concentrations, the method is accurate in the hands of a skillful technician, but the results may not be reproducible, and it does not lend itself to an estimate of sampling error (Cohn, 1995). Since mean daily concentration must be estimated from instantaneous concentration, errors may be introduced for constituents that vary widely over the course of a day.

Ratio estimators assume a constant ratio between two variables, usually concentration and discharge, or load and discharge (Cohn, 1995). A ratio estimator is a best linear unbiased estimator provided that: (a) samples are collected at random; (b) the relation

between y_i and x_i is a straight line through the origin; and (c) the variance of y_i about this line is proportional to x_i , where y_i is the dependent variable and x_i is the independent variable. This condition is often met with instantaneous load as the dependent variable and instantaneous discharge as the independent variable (Preston *et al.*, 1989). An example of this approach is the Beale's Ratio Estimator (BRE, described in detail below).

Regression estimators have long been used to estimate loads of suspended sediment, usually in a log-log form, since both concentration and discharge are assumed to be log-normally distributed. Log of instantaneous concentration (C_i) is regressed against log of instantaneous discharge (Q_i), and the resulting relationship can be used to predict daily concentration (C_d) from daily discharge (Q_d), provided that a correction factor for retransformation bias is introduced (Ferguson, 1986; Cohn, 1995).

A variant of the rating curve method has been used by the University of California-Davis Tahoe Research Group (TRG) to calculate total nutrient loads for the Tahoe Basin (Byron *et al.*, 1989). Instead of $\log C_i$ vs. $\log Q_i$, instantaneous load (L_i) is calculated as the product $C_i \cdot Q_i$, and regressed against $\log Q_i$. The resulting relationship (with appropriate correction for retransformation bias) is used to estimate L_d from Q_d , and the estimates are summed over days for the water year. The load estimates by this variant, however, are mathematically identical to those obtained by a regression of $\log C_i$ vs. $\log Q_i$. The apparent high correlation between $\log L_i$ and $\log Q_i$ is a "spurious self-correlation" (Galat, 1990).

Stratification of discharge and concentration data by flow class, month or season can appreciably improve the accuracy of load estimates (Richards and Holloway, 1987; Preston *et al.*, 1989). It can be applied to any of the main load calculation methods. Hill (1986), for example, developed separate nitrate-N rating curves for the November to April and May to October periods for Ontario streams, where nitrate-N concentrations are typically two orders of magnitude higher than in Tahoe Basin streams. In a Monte Carlo study of load calculation methods for calculating TP loads in a tributary of the Great Lakes, Dolan *et al.* (1981) showed that applying the BRE separately to just two flow classes significantly improved the method's accuracy.

Thomas (1985) developed a variable-probability sampling method for suspended sediment load estimation, in which the probability of collecting a sample is proportional to its estimated contribution to total suspended sediment discharge. This method was later compared with time-stratified sampling and flow-stratified sampling (Thomas and Lewis, 1993). Such sampling designs allow for unbiased estimates of total

load, as well as estimation of sampling error; their implementation may be facilitated with automated and preprogrammed sampling equipment.

Using data for three large river basins in Ohio, Richards and Holloway (1987) simulated concentrations at six-hour intervals for nitrate, total phosphorus (TP), soluble reactive phosphorus (SRP), suspended solids (SS), and conductivity for 1,000 years. They then sampled the synthetic data sets to evaluate both sampling and load calculation methods. They found that the bias and precision of load estimates are affected by frequency and pattern of sampling, calculation approach, watershed size, and the behavior of the chemical species being monitored. Of the strategies evaluated, the BRE with flow-stratified sampling provided the best results.

The purpose of this study was to evaluate and compare four alternative methods for estimating tributary mass loads of different forms of nitrogen and phosphorus, and of suspended sediment, for the given sampling program of the LTIMP. It was hoped that the results would: (1) provide guidance to the LTIMP in the choice of methods for estimating total loads from existing data; (2) provide guidance on possible modification to the sampling program; and (3) provide insights that may be useful to other monitoring programs that are attempting to measure nutrient loads in mountain streams.

METHODS

It is common in water quality sampling programs for samples to be collected nonrandomly. This complicates the estimation of the bias and precision of load estimates. A Monte Carlo test in which a large data set is resampled many times is one way around the problem of lack of random sampling.

In the Monte Carlo test, we resampled water quality data sets (with replacement) from 1985 (89 samples) and 1986 (136 samples) for Blackwood Creek, a major tributary in the Tahoe basin. For each test we created 200 data subsets with samples sizes (n) of 20, 40, 60, and 80 samples for both years, and additionally of 100 and 120 samples for 1986. We thus had 800 data subsets for 1985 and 1,200 subsets for 1986. These subsets were then used to calculate total annual loads of total phosphorus (TP), particulate phosphorus (PP), soluble reactive phosphorus (SRP), nitrate-nitrogen and suspended sediment (SS), by four methods, as described below. These load estimates were then compared with load estimates derived by the worked record method.

The Blackwood water quality samples were collected by the Lake Tahoe Interagency Monitoring Program (LTIMP). The LTIMP stream sampling network was setup in 1979, to monitor seven tributary streams for nutrients and sediment. By 1987, the network had been cut back to four streams. It was then expanded to include four small tributaries on the Nevada side. By 1993, the network included 32 sites in 14 basins, with a total of 20 stream gaging stations (Boughton *et al.*, 1997). Figure 1 shows the tributaries sampled in the program.

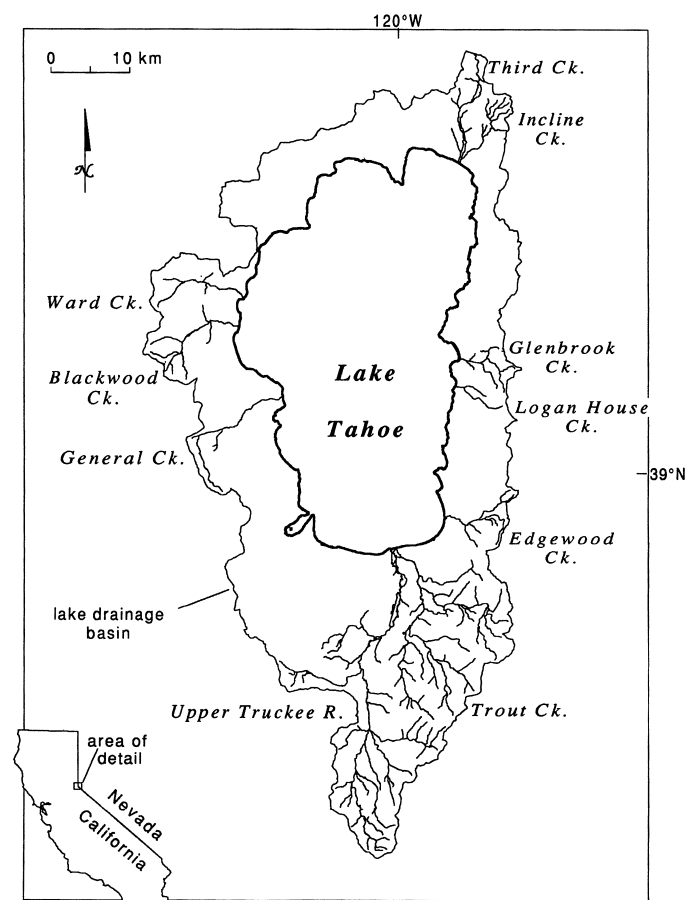


Figure 1. The Lake Tahoe Basin, Showing the LTIMP Sampling Tributaries.

Samples for nutrient and sediment analysis are collected and handled according to protocols established by the U.S. Geological Survey, which maintains the gaging stations and discharge records. Throughout the program, stream sampling has been concentrated during periods of high flow; sampling crews often work at night, and on weekends and holidays. Until the mid-1990s, typically greater than 100 samples were collected per year at the major stations

at tributary mouths. During spring snowmelt, samples were usually collected at least daily at the major stations. Since the mid-1990s, the sampling intensity has been reduced, but the number of stations has been maintained. In WY 1998, the LTIMP collected an average of 29 samples at each of the ten tributary mouth stations.

Samples are stored on ice and sent either to a laboratory in the Tahoe basin, at the University of California-Davis or (for suspended sediment) to the USGS laboratory in Salinas, California. Constituents analyzed include nitrate-nitrogen, ammonia-nitrogen, Total Kjeldahl Nitrogen (TKN, since 1989), soluble reactive phosphorus (SRP), total phosphorus (TP), biologically-available iron, and suspended sediment. In some years, a limited set of samples has been filtered (0.45 μm), allowing calculation of particulate (PP) and dissolved (DP) phosphorus, and particulate (PON) and dissolved organic nitrogen (DON) by difference. The chemical methods have been described in detail elsewhere (Goldman *et al.*, 1993). Quality assurance/quality control procedures (use of field blanks, spike recovery, duplicate samples, etc.) conform to USGS protocols.

Blackwood Creek, the source of our data, drains an area of 29 km² on the west side of the Lake Tahoe basin. Its mean annual runoff is 114 cm, most of which occurs during spring snowmelt. There is a wide range in discharge over the water year; the mean annual maximum daily discharge is 342 times the mean annual minimum daily discharge. There is little development in the basin, but it was subjected to heavy logging and grazing up to the early 1960s. The annual runoff for Blackwood Creek in 1985 was 30 percent below the mean for 1961 to 1998, and in 1986 it was 45 percent above the long-term mean. These two water years thus represent a wide range of runoff conditions.

The mean daily discharge, measured concentrations and worked record concentration estimates of TP, PP, and SS are shown in Figure 2a; SRP and nitrate-N are shown in Figure 2b. The points represent the sample concentrations that were resampled in the Monte Carlo experiment. Ammonia-N data were not used, since the concentrations of that form are generally very low in basin streams and contribute a negligible amount to total nitrogen load (Coats and Goldman, 2001). Table 1 shows the coefficient of variation (standard deviation as percent of mean) of instantaneous loads sampled by the LTIMP in 1985 and 1986.

For these two years, the TRG staff had estimated the average daily concentration, by the "worked record" method, for nitrate-N, SRP, total hydrolyzable phosphorus (THP), and suspended sediment. These

average daily concentration estimates were multiplied by the mean daily discharge (from the USGS record) to give the average daily loads, which were then summed over each water year. The resulting estimates of total annual load are taken as the "true" loads against which the estimates by the four test methods are compared. We recognize that the worked record method may introduce its own bias, especially for particulate constituents, but at this point it seems to provide the best available standard against which to judge estimates based on a smaller sample size. For nitrate-N, a discharge-concentration model may also be used to generate a trace of C_d (Coats and Goldman, 2001), but for the other constituents, we do not have a model for simulating C_d or C_i that would not incorporate the same assumptions as the load calculation models we are trying to test.

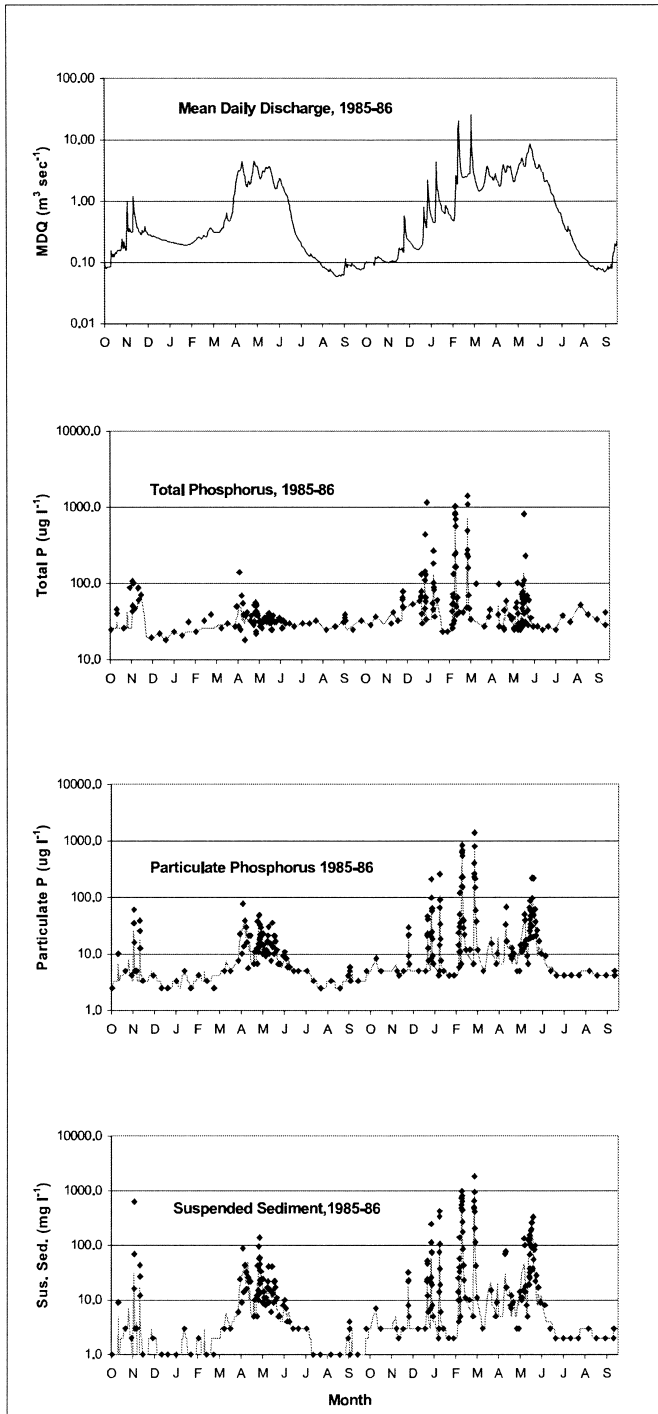
The THP analysis, since replaced with a different digestion procedure, gives results that are somewhat less than but linearly related to total phosphorus (TP). To maintain comparability with the current methods, the 1985 and 1986 THP data and worked record estimates were converted by regression to TP (Hatch, 1997). Particulate phosphorus (PP) for Blackwood Creek was estimated by regression with suspended sediment ($R^2 = 0.85$, $n = 124$), but constrained so that the PP value for each water sample in the data sets never exceeded (TP)-(SRP).

The sampling program of the LTIMP does not use a rigorous random or periodic sampling scheme; rather sampling decisions are based on the professional judgment of field crews, who must deal with access during severe storms as well as budget limitations. We assumed that the bias toward high discharge that is built into the LTIMP sampling effort would be preserved in a random subsample of the original data. The sampling was constrained by the requirement that each annual quartile of mean daily discharge be represented by at least three samples in each data set. This is equivalent to assuming that the field crews will sample a range of discharges, even though their efforts are concentrated at high discharge. It is possible, however, that the field crews could do a better job of distributing their sampling efforts over flow class than that represented by our constrained random sampling.

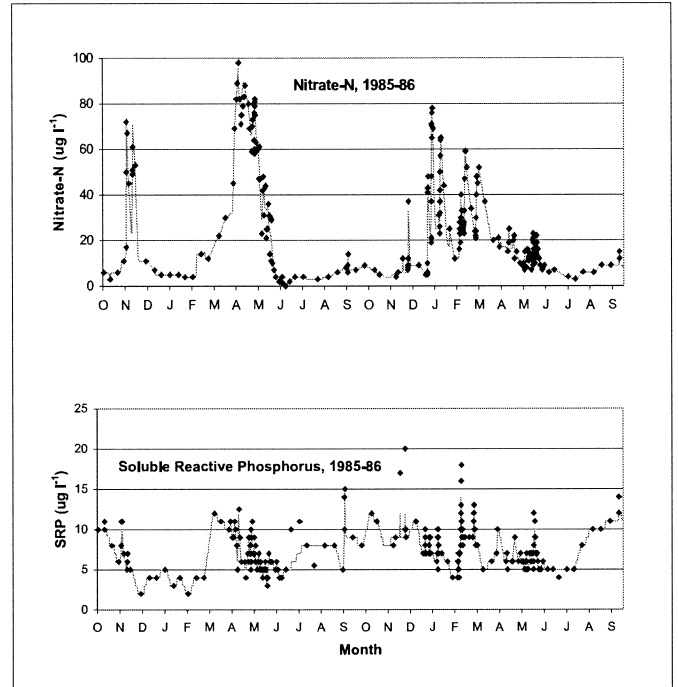
Using the 800 data subsets for 1985 and 1,200 subsets for 1986, we then calculated the total load for the two water years, by four different methods.

The Beale's Ratio Estimator (BRE)

In this method, the discharge-weighted mean concentration is multiplied by total discharge in the



(a)



(b)

Figure 2. Mean Daily Discharge, and Constituent Concentrations, Blackwood Creek, Water Years 1985 and 1986. Figure 2a shows mean daily discharge, total P, particulate P, and suspended sediment; Figure 2b shows nitrate-N and soluble reactive P. Points represent the actual water quality samples (used in the Monte Carlo tests), and the lines represent the worked record estimates of mean daily concentration.

TABLE 1. Coefficients of Variation of Instantaneous Load, for the Data Sets Used in the Monte Carlo Test.

	C.V. (percent)	
	1985	1986
Nitrate-N	103	114
Soluble Reactive P	82	176
Particulate P	119	409
Total P	86	379
Suspended Sediment	251	383

defined time interval, and the result adjusted using a factor that incorporates the ratio of the covariance of load with discharge to the variance of discharge. The BRE was chosen for this study because it has been used successfully to estimate total phosphorus loads in other areas.

The equation used from Cohn (1995) is

$$\bar{L} = \bar{Q} \cdot \frac{\bar{L}^*}{\bar{Q}^*} \cdot \left(\frac{1 + \frac{S_{LQ}}{N \cdot \bar{L}^* \cdot \bar{Q}^*}}{1 + \frac{S_{Q^2}}{N \cdot \bar{Q}^{*2}}} \right) \quad (2)$$

where

\bar{Q} = the average water discharge during the defined time period;

L_i = the instantaneous load $Q_i C_i$

$$\bar{L}^* = \frac{1}{N} \sum_{i=1}^N L_i$$

$$\bar{Q}^* = \frac{1}{N} \sum_{i=1}^N Q_i$$

$$S_{LQ} = \frac{1}{N-1} \sum_{i=1}^N (Q_i - \bar{Q}^*) \cdot (L_i - \bar{L}^*)$$

$$S_{Q^2} = \frac{1}{N-1} \sum_{i=1}^N (Q_i - \bar{Q}^*)^2; \text{ and}$$

N = the number of samples in the defined time period

The Stratified Beale's Ratio Estimator (SBRE) Method

We stratified the data (*a posteriori*) according to the four quartiles of mean daily discharge, calculated the load separately for each quartile, and summed the results.

The Period Weighted Sample Method

In this method, the concentration in each two successive samples in the water year were averaged and the mean multiplied by the cumulative discharge between sampling times. Discharge was taken from

the USGS record of mean daily discharge, subdivided for days on which one or more samples were taken. The resulting increments of load were then summed over each of the water years.

The Rating Curve Method

The log of instantaneous concentration ($\log C_i$) was regressed against log of instantaneous discharge ($\log Q_i$). For each day in the water year, mean daily discharge was then used to estimate mean daily concentration, according to the equation:

$$C_d = k \cdot 10^a \cdot Q_d^b \cdot e^{2.65 \cdot \text{MSE}} \quad (3)$$

where C_d is the mean daily concentration; k is a unit correction factor; a is the regression constant; b is the regression coefficient; and MSE is the mean square error from the regression of $\log C_i$ vs. $\log Q_i$.

This is essentially the method used by the Tahoe Research Group to calculate total constituent loads in basin streams, except that the regression they use is $\log(Q_i C_i)$ vs. $\log Q_i$. The factor $e^{2.65 \text{MSE}}$ is a correction factor for retransformation bias (Ferguson, 1986).

Using the 200 total load estimates for each method, year and sample size, we calculated means, standard deviations, and expressions for bias and imprecision. Bias for a method is defined as the deviation of the mean of load estimates from the worked record estimate (as percent of the latter); imprecision is defined as the Coefficient of Variation of a method (standard deviation as percent of method mean). We also calculated the standard deviations as percent of the worked record estimates.

A useful statistic for expressing both the bias and imprecision of an estimate is the root mean square error (RMSE), defined as

$$\text{RMSE} = \sqrt{B^2 + S^2} \quad (4)$$

where B is the deviation from the worked record estimate, and S is the standard deviation of the sample (Dolan *et al.*, 1981). This statistic was calculated for all constituents, methods and sample sizes.

RESULTS

Table 2 shows the results of the Monte Carlo experiment for $n = 40$, a somewhat larger sample size than presently used by the LTIMP. The standard deviation for each set of 200 load calculations is shown as both

TABLE 2. Imprecision and Bias of Four Load Calculation Methods, From a Monte Carlo Test of 40 Samples Drawn 200 Times With Replacement From Water Quality Data Sets of 89 Samples (1985) and 136 Samples (1986).

	1985				1986			
	S.D. as Percent Method Mean	S.D. as Percent Worked Record	Bias as Percent Worked Record	RMSE as Percent Worked Record	S.D. as Percent Method Mean	S.D. as Percent Worked Record	Bias as Percent Worked Record	RMSE as Percent Worked Record
Nitrate-N								
BRE	9.5	11.9	25.0	27.7	8.0	8.9	12.0	15.0
SBRE	8.8	9.9	11.3	15.4	8.6	10.2	18.3	20.9
PWS	4.3	4.2	-2.3	4.8	8.3	8.3	-0.6	8.3
RC	10.8	14.4	33.3	36.3	12.6	14.7	16.4	22.0
Soluble Reactive P								
BRE	5.5	5.7	3.0	6.4	9.5	11.6	21.8	24.7
SBRE	5.1	5.4	4.6	7.0	7.7	9.2	19.3	21.4
PWS	5.1	5.2	1.3	5.3	5.5	5.7	3.0	6.5
RC	4.8	5.2	8.7	10.2	6.0	6.5	7.1	9.6
Particulate P								
BRE	11.1	16.8	51.8	54.4	48.6	161	231	282
SBRE	10.5	13.8	31.2	34.1	47.2	132	178	222
PWS	14.1	15.1	7.3	16.8	48.9	56.2	14.9	58.1
RC	10.3	12.5	21.3	24.7	23.5	19.4	-17.6	26.2
Total P								
BRE	7.2	8.6	20.4	22.1	43.6	138	216	256
SBRE	6.7	8.1	19.8	21.4	40.9	112	174	207
PWS	7.4	8.0	8.7	11.9	38.6	48.6	26.0	88.1
RC	6.8	8.3	21.5	23.0	22.8	24.2	7.5	25.6
Suspended Sediment								
BRE	30.0	74.3	150	168	43.4	149	244	286
SBRE	29.1	61.0	109	125	42.3	121	186	222
PWS	23.5	28.3	20.2	34.8	43.5	53.8	23.6	58.8
RC	21.4	33.3	55.3	64.6	22.8	22.2	-3.1	22.4

percent of the method mean (Coefficient of Variation) and as percent of the worked record estimate. The bias for each method is shown as the departure of the mean for each set of load calculations from the worked record estimate (as percent of the latter). Table 2 also shows the RMSE for each constituent and method, combining the effect of both bias and imprecision. Table 3 summarizes the regression results for the rating curve method, for n = 40.

For nitrate-nitrogen, the Period Weighted Sample (PWS) method is clearly superior to the other methods, both in its precision and lack of bias. For 1985, the SBRE was less biased than the BRE, but for 1986, it was more biased. The Rating Curve (RC) method was biased and imprecise in both years, with higher

RMSE than the BRE or SBRE. Note that in 1986 (the wet year), only 66.5 percent of the nitrate-N regressions were significant at the 95 percent level, with an average R² of 0.14 (Table 3).

For SRP, all of the methods appeared to work reasonably well in 1985, but the Beale's Ratio Estimators were highly biased (on the high side) in 1986. The PWS method performed about the same as the rating curve method in terms of precision, but was superior in terms of bias. Note in Table 3, however, that the rating curve regressions for SRP were mostly insignificant.

For TP, the PWS was the least biased of the methods in 1985, but slightly less precise than the other methods. The RC estimates in 1985, however, are

TABLE 3. Summary of Rating Curve Regression Results, for 200 Regressions of Log C_i vs. Log Q_i , With $n = 40$.

	1985		1986	
	Average R^2	Percent of Regressions Significant at 95 Percent Level	Average R^2	Percent of Regressions Significant at 95 Percent Level
Nitrate-N	0.33	98	0.14	66.5
Soluble Reactive P	0.05	15	0.03	8.5
Particulate P	0.52	100	0.58	100
Total P	0.03	3.5	0.25	91.5
Suspended Sediment	0.56	100	0.64	100

based on regressions that were virtually all nonsignificant (Table 3). In 1986, the RC method was the least biased and the most precise of the four methods, with a CV of 22.8 percent, (at $n = 40$; see Table 2) departure from the worked record estimate of 7.5 percent, and the lowest RMSE of the four methods. Most (greater than 90 percent) of the TP rating curve regressions were significant for 1986; although the average R^2 for the 200 regressions was only 0.25. The BREs again performed poorly, especially in 1986. Stratification slightly improved the performance of the BRE, but not enough to justify its use for the Tahoe basin.

For suspended sediment, the BREs were both highly biased and highly imprecise. The PWS was also biased and imprecise, for both years, although it was less biased than the rating curve method for 1985. For 1986, the rating curve method was the most precise of the methods tested for suspended sediment (CV = 22.8 percent). With a deviation from the worked record of -3.1 percent, it was the least biased of the methods tested.

Since particulate P (PP) was estimated by regression (constrained at the high end; see Methods) with suspended sediment (SS), the results for PP similar but not identical to the results for SS. The PWS method gave the least biased results for 1985, but was slightly less precise than the rating curve results. For 1986, the rating curve method was the least biased and the most precise.

In general, the CV of instantaneous load (Table 1) was a good predictor of the imprecision of the load estimates, especially for 1986. The particulate constituents (SS, TP, and PP) had the highest CVs of instantaneous load, and the most imprecise load estimates. The dissolved constituents (nitrate-N and SRP) had less variable instantaneous loads, and consequently, the total load estimates are more precise. This is consistent with the results of Richards and

Holloway (1987). Note, however, that the CVs represent the distribution of load estimates based on sets of subsamples drawn from finite populations of actual samples. The true CVs of load estimates drawn from the continuously-varying concentrations are unknown, but no doubt greater.

Figure 3 shows the relationship between sample size and RMSE, for the four methods, five constituents and two water years. In 1985, the PWS method had the lowest RMSE of the four methods, for all constituents and sample sizes. In 1986 the RC method outperformed the PWS method for PP, TP, and SS but the advantage diminished with sample size, since the performance of the PWS method improves with sample size more than that of the other methods. This is to be expected, since a PWS calculation using all of the data is similar to the worked record calculation. Since the subsample sets comprise a significant fraction of the data sets from which they were drawn, the RMSEs overstate the improvement of accuracy with sample size. They are, however, useful for comparing load calculation methods for a given sample size.

DISCUSSION

The choice of a method for calculating total load depends on: (1) the physical and biological processes that control the discharge-concentration relationship for a constituent; (2) the statistical distribution of a constituent; and (3) the sampling regime. An appropriate method for one constituent may be inappropriate for another. Since an important objective of the sampling program is estimating total loads over time, the results for heavy runoff years (e.g., 1986) should be given more consideration than test results for a dry year. It would be useful, of course, to know the fre-

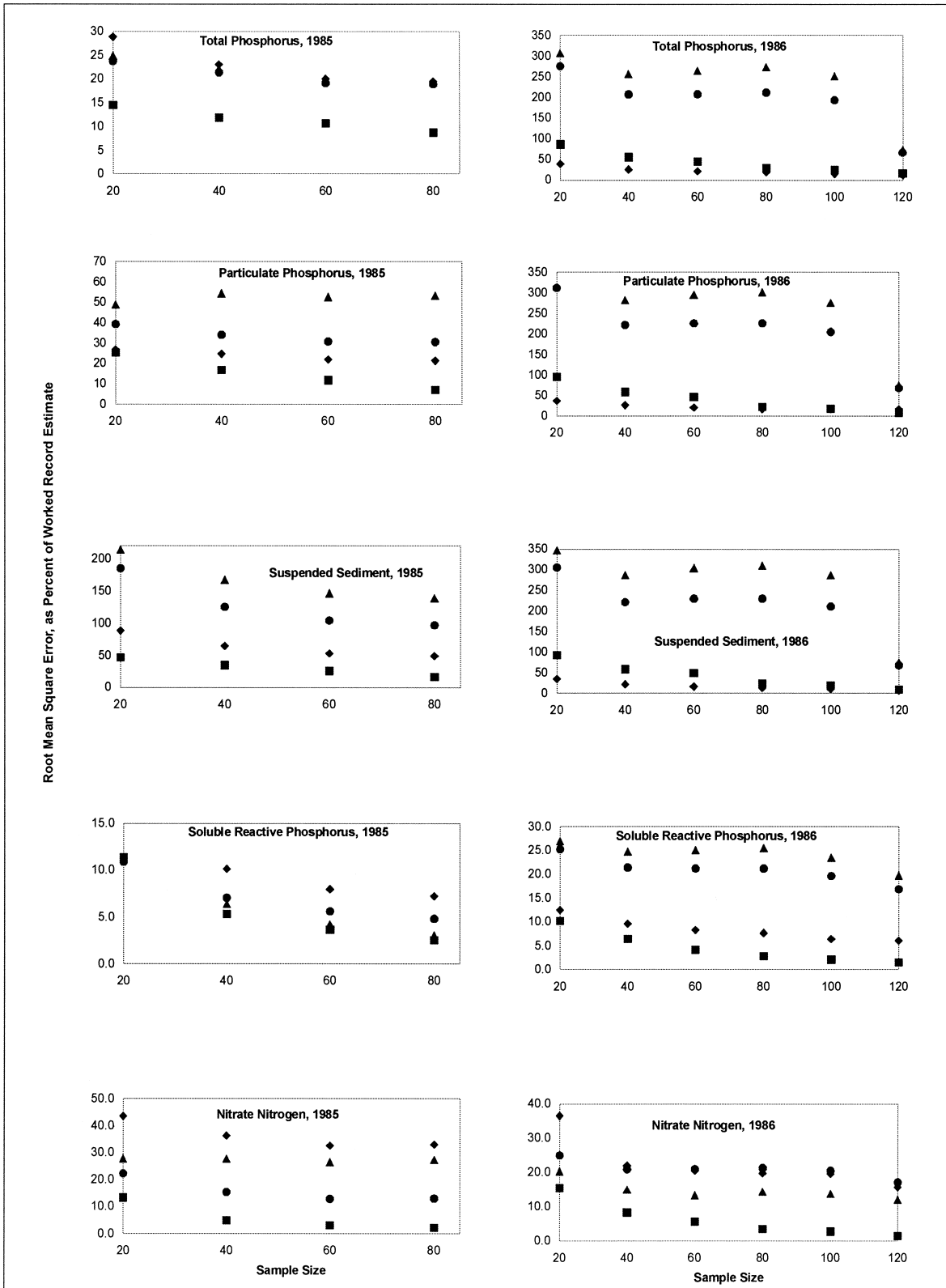


Figure 3. Relationships Between Sample Size and Root Mean Square Error for Total Phosphorus, Particulate Phosphorus, Suspended Sediment, Soluble Reactive Phosphorus, and Nitrate-Nitrogen. ◆ Rating Curve; ■ Period-Weighted Sample; ▲ Beale's Ratio Estimator; and ● Stratified Beale's Ratio Estimator.

quency of the annual runoff that accounts for most of the load in the long run, and compare methods for such a year.

Nitrate-nitrogen concentrations in basin streams are influenced by microbial release, plant uptake (in both the terrestrial and lotic systems), precipitation and snowmelt input, and washout during winter rainstorms and early snowmelt (Coats and Goldman, 2001). The discharge-concentration relationship in basin streams typically shifts from fall to late spring. The sampling program is designed to sample high flows more intensively than low flows, but there is no assurance that sampling frequency is proportional to discharge. Since high flows during the snowmelt recession period are typically very low in nitrate-N and are not sampled as intensively as the early snowmelt period, there is a possibility of bias toward high concentrations, and the Rating Curve results seem to reflect this. We recommend the PWS method for calculating nitrate-N loads. This conclusion is consistent with a comparison of the PWS with RC methods on the basis of simulated daily nitrate-N concentration (Coats and Goldman, 2001).

Load estimates based on the PWS for basin streams, however, have shown that nitrate-N accounts for only 10 to 15 percent of the nitrogen load. Organic-N (about 55 percent of which is dissolved, on average), accounts for almost all of the rest; ammonium loads are negligible (Coats and Goldman, 2001). It would be useful to run a similar Monte Carlo test for organic-N, but it was not measured effectively until 1989. Like nitrate-N, however, it is strongly influenced by microbial release and wash-out during fall storms and early snowmelt. In some basin streams, dissolved organic nitrogen (DON) increases again at low-flow, probably due to periphyton growth and decay. We consider the PWS method to be tentatively appropriate for organic-N as well as nitrate-N, but since organic-N has an important particulate component, the rating curve method may deserve more consideration.

Soluble reactive phosphorus (SRP) is strongly controlled by adsorption onto fine sediment particles, mostly oxides of Fe and Al (Froelich, 1988). Although the sorption-desorption process is complex, it is apparently rapid enough that the SRP concentration is virtually independent of discharge. Both the PWS and Rating Curve methods produced satisfactory results. Even the Beale's Ratio Estimators were satisfactory for 1985, but not for 1986. Most of the rating curve regressions, however, were not statistically significant (Table 2). There is no theoretical justification for using the Rating Curve method for SRP, which bases regression predictions on statistically-insignificant relationships. Of course, if L_i is used as the dependent variable and Q_i as the independent

variable, the regressions will look better, but the results – and lack of theoretical justification, in the case of SRP – are the same. We recommend using the PWS method for SRP.

Our results for SRP contrast with those of Johnson (1979), who found a strong linear relationship between discharge and SRP concentration in Fall Creek, New York, with concentrations as high as 80 $\mu\text{g/l}$. The linear rating curve was found in that case to give a better load estimate than averaging methods using sampling proportional to discharge or flow duration. The best estimate, however, was obtained by a multiple regression of concentration versus discharge (Q) and rate of change of discharge (dQ/dt).

In Tahoe basin streams, most of the total phosphorus load is bound to particles, either by adsorption or as part of an organic particle or primary mineral. The size of the particles has a strong influence on the way in which particulate phosphorus (PP) responds to changes in discharge. Hatch *et al.* (1999) measured PP separately for the sand fraction (greater than 0.63 μm) and silt plus clay fraction (greater than 0.45 μm , less than 0.63 μm) during spring runoff in Incline Creek, on the northeast side of the Tahoe basin. They found that the sand-sized PP was depleted during the daily snowmelt flood, but that the silt plus clay PP concentration increased with falling discharge, suggesting a groundwater (or shallow soil-water) source. Note that in Blackwood Creek, the TP and PP concentration fluctuated widely and rapidly, (Figure 2a) especially in 1986, when the CVs of instantaneous load were 379 and 409 percent, respectively.

Since primary productivity in the lake is apparently phosphorus-limited, the choice of a method for estimating TP loads is important. For TP, the RMSE in 1985 was lowest for the PWS method, and in 1986 for the RC method. Based on these results, we recommend using the rating curve method for years in which the $\log C_i$ versus $\log Q_i$ regression for TP is statistically significant. For years in which it is not, the PWS method should be used.

Our results for TP contrast somewhat with those of Dolan *et al.* (1981). Using a similar Monte Carlo approach, those authors compared ratio estimator and regression methods for estimating TP loads in a large Michigan river basin with stable flow and large fertilizer inputs of phosphorus. They found that the SBRE (with two flow strata) was superior to the regression methods tested. They noted, however, that the ratio estimator methods “would probably not be appropriate for excessively variable streams in which daily flows vary by a factor of 100 or more.” We agree.

For suspended sediment (SS), the Rating Curve method was remarkably unbiased for 1986, but over-predicted by 55 percent for 1985. Since the estimated SS loads were more than an order of magnitude

greater in 1986 than in 1985, the Rating Curve method is a better choice than the PWS. This is fortunate, since this method is widely used and accepted for estimating SS loads. Recent experiments with continuous turbidity probes in Ward Creek (see Figure 1) have shown that they offer considerable promise for measuring suspended sediment loads in basin streams (A. Stubblefield, pers. comm., 2001).

For large river basins (386 to 16,700 km²) draining to Lake Erie, Richards and Holloway (1987) found that the Beale's Ratio Estimator with flow-stratified sampling gave the best results, for all constituents. They cautioned that the flux variance over time in small rivers is greater relative to mean flux than in large rivers, and concluded that for their large rivers, "... programs employing infrequent sampling (less than about 50 samples/year) will generally provide load estimates which are strongly biased and very imprecise ... (such) programs may produce load estimates which are subject to such great uncertainty as to be of questionable utility for water resource management." The many small tributaries in the Lake Tahoe basin, with high annual variance of daily flow, present a difficult sampling challenge for the Lake Tahoe Interagency Monitoring Program.

The rating curve method for TP in 1986 was relatively unbiased; with 40 samples, the deviation of the average of the 200 calculations from the worked record was only 7.5 percent. For some purposes, however, the precision of the method is more important than lack of bias in calculating total tributary loads. For example, investigators might want to compare TP loads for different tributaries, and relate them to soils, hydrology, geomorphology, and land use, or measure the effectiveness of nutrient control programs. A consistent bias might not affect the ultimate conclusions if tributary differences are being examined. But poor precision would make it difficult to compare tributary loads. With 30 samples for 1986, the CV for Blackwood Creek's TP load (of about 190 kg km⁻²) by the Rating Curve method would be about 30 percent, and the 95 percent confidence limits would be ± 60 percent of the load estimate. If a watershed were treated to reduce the tributary load of TP and monitored for a single year with 30 samples each from the treatment and control watersheds, (and the same CV assumed) the difference would have to be at least 74 percent of the estimated load from the control watershed in order for the difference to be considered significant at the 95 percent level. A pretreatment and longer-term data set would improve the power of the test, but clearly the high variance of TP load estimates is a problem for measuring the effectiveness of control programs.

SUMMARY AND CONCLUSIONS

This Monte Carlo experiment has shown that the Period Weighted Sample method is superior to both the Rating Curve and unstratified Beale's Ratio Estimator, for estimating nitrate-N loads. It is somewhat less biased than the other two methods for SRP, whereas the Rating Curve method for SRP relies on statistically-insignificant regressions. For TP, the Period Weighted Sample method was the best for the dry year of 1985, and the Rating Curve method was best for the wet year of 1986. The Rating Curve method performed well for SS in a heavy runoff year, but not in a dry year. In general, the Period Weighted Sample method performed best for dissolved constituents and the rating curve method performed best for particulate constituents.

Additional evaluation of methods for calculating total loads in Tahoe basin streams is needed. New studies might include: (1) comparing methods of load calculation for organic nitrogen (which would require developing some "worked record" estimates for one or two years); (2) subdividing the data by season and flow regime, and testing rating curves that take account of seasonal and event hysteresis; (3) developing a statistically-based flow-stratified or time-stratified sampling program, with random sampling within strata. Such an improved sampling program might also incorporate continuous turbidity probes.

The U.S. Congress, along with California and Nevada, recently appropriated \$600 million to restore and protect the environment in the Tahoe Basin. Support for this bill was generated largely by concern for the lake's diminishing clarity, and much of the money will be spent on projects to control the flux of phosphorus to the lake. The results of this study suggest that in order to measure the effectiveness of these projects, the sampling frequency and data analysis methods used in the Tahoe Basin will have to be modified.

ACKNOWLEDGMENTS

We thank David Dawdy, John Reuter, Bob Thomas and four anonymous reviewers for helpful suggestions and critical review of the manuscript. Special thanks are due Mark Williams for his support and patience. The data used in this study were collected by the Lake Tahoe Interagency Monitoring Program, which is supported by the California State Water Resources Control Board, the U.S. Geological Survey, the Tahoe Regional Planning Agency, the Lahontan Regional Water Quality Control Board, the University of California, the California Department of Fish and Game, the California Department of Water Resources, the California Department of Transportation, the California Air Resources Board, the Nevada Department of Environmental Protection, the U.S. Environmental Protection Agency, and the U.S. Forest Service.

LITERATURE CITED

- Boughton, D. J., T. G. Rowe, K. K. Allander, and A. R. Robledo, 1997. Stream and Ground-Water Monitoring Program, Lake Tahoe Basin, Nevada and California. FS-100-97, U.S. Geological Survey, Carson City, Nevada, 6 pp.
- Byron, E. R., C. R. Goldman, and S. H. Hackley, 1989. Lake Tahoe Interagency Monitoring Program Ninth Annual Report, Water Year 1988. Tahoe Research Group Institute of Ecology. Univ. of Calif. at Davis, Davis, California, 62 pp.
- Chang, C. C. Y., J. S. Kuwabara, and S. P. Pasilis, 1992. Phosphate and Iron Limitation of Phytoplankton Biomass in Lake Tahoe. *Canadian Journal of Fisheries and Aquatic Sciences* 49(6):1206-1215.
- Coats, R. N. and C. R. Goldman, 2001. Patterns of Nitrogen Transport in Streams of the Lake Tahoe Basin, California-Nevada. *Water Resources Research* 37(2):405-416.
- Cohn, T. A., 1995. Recent Advances in Statistical Methods for the Estimation of Sediment and Nutrient Transport in Rivers. *Reviews of Geophysics, Supplement*: 1117-1123.
- Dann, M. S., J. A. Lynch, and E. S. Corbett, 1986. Comparison of Methods for Estimating Sulfate Export From a Forested Watershed. *Journal of Environmental Quality* 15(2):140-145.
- Dolan, D. M., A. K. Yui, and R. D. Geist, 1981. Evaluation of River Load Estimation Methods of Total Phosphorus. *Journal of Great Lakes Research* 7(3):207-214.
- Ferguson, R. I., 1986. River Loads Underestimated by Rating Curves. *Water Resources Research* 22(1):74-76.
- Ferguson, R. I., 1987. Accuracy and Precision of Methods for Estimating River Loads. *Earth Surface Processes and Landforms* 12(1):95-104.
- Froelich, P. N., 1988. Kinetic Control of Dissolved Phosphate in Natural Rivers and Estuaries: A Primer on the Phosphate Buffer Mechanism. *Limnology and Oceanography* 33:649-668.
- Galat, D. L., 1990. Estimating Fluvial Mass Transport to Lakes and Reservoirs: Avoiding Spurious Self-Correlations. *Lake and Reservoir Management* 6(2):153-163.
- Goldman, C. R., 2000. Baldi Lecture. Four Decades of Change in Two Subalpine Lakes. *Verhandlungen der Internationale Vereinigung für Theoretische und Angewandte Limnologie* 27:7-26.
- Goldman, C. R., A. D. Jassby, and S. H. Hackley, 1993. Decadal, Interannual, and Seasonal Variability in Enrichment Bioassays at Lake Tahoe, California-Nevada, USA. *Canadian Journal of Fisheries and Aquatic Sciences* 50:1489-1495.
- Hatch, L. K., 1997. The Generation, Transport, and Fate of Phosphorus in the Lake Tahoe Ecosystem. Ph.D. thesis, Univ. of Calif., Davis, California, 212 pp.
- Hatch, L. K., J. E. Reuter, and C. R. Goldman, 1999. Daily Phosphorus Variation in a Mountain Stream. *Water Resources Research* 35(12):3783-3791.
- Hill, A. R., 1986. Stream Nitrate-N Loads in Relation to Variations in Annual and Seasonal Runoff Regimes. *Water Resources Bulletin* 22(5):829-839.
- Jassby, A. D., C. R. Goldman, and J. E. Reuter, 1995. Long-Term Change in Lake Tahoe (California-Nevada, U.S.A.) and Its Relation to Atmospheric Deposition of Algal Nutrients. *Archive für Hydrobiologie* 135:1-21.
- Johnson, A. H., 1979. Estimating Solute Transport in Streams From Grab Samples. *Water Resources Research* 15(5):1224-1228.
- Likens, G. E., F. E. Bormann, R. S. Pierce, J. S. Eaton, and N. M. Johnson, 1977. *Biogeochemistry of a Forested Ecosystem*. Springer-Verlag, New York, New York, 159 pp.
- Preston, S. D., V. J. Bierman, Jr., and S. E. Silliman, 1989. An Evaluation of Methods for the Estimation of Tributary Mass Loads. *Water Resources Research* 25(6):1379-1389.
- Reuter, J. E., C. R. Goldman, T. A. Cahill, S. S. Cliff, A. C. Heyvaert, A. D. Jassby, S. Lindstrom, and D. M. Rizzo, 1999. An Integrated Watershed Approach to Studying Ecosystem Health at Lake Tahoe, California-Nevada, USA, International Congress on Ecosystem Health, Sacramento, California (in press).
- Richards, R. P. and J. Holloway, 1987. Monte Carlo Studies of Sampling Strategies for Estimating Tributary Loads. *Water Resources Research* 23(10):1939-1948.
- Thomas, R. B., 1985. Estimating Total Suspended Sediment Yield with Probability Sampling. *Water Resources Research* 21(9):1381-1388.
- Thomas, R. B. and J. Lewis, 1993. A Comparison of Selection at List Time and Time-Stratified Sampling for Estimating Suspended Sediment Loads. *Water Resources Research* 29(4):1247-1256.