Hydroikos Ltd.

2512 Ninth Street, Ste. 7
Berkeley, CA 94710
Phone: (510) 295-4094
coats@hydroikos.com
www.hydroikos.com

REALIGNING THE LAKE TAHOE INTERAGENCY MONITORING PROGRAM
Vol. I: Main Report

By

Robert Coats and Jack Lewis

July 2014

Table of Contents

List of Figures

List of Tables

List of Appendices

List of Acronyms and Abbreviations

ac	Acres
ACE	Alternating Conditional Expectations (see Glossary)
AIC	Akaike's Information Criterion (see Glossary)
AVAS	Additivity and Variance Stabilization for regression (see Glossary)
BAP	Biologically Available Phosphorus
BMP	Best Management Practice
^oC	Degrees Celsius
cfs	cubic feet per second
CWA	Clean Water Act
DCNR	Nevada Department of Conservation and Natural Resources
DIN	Dissolved Inorganic Nitrogen
DON	Dissolved Organic Nitrogen
DRI	Desert Research Institute
Et	Evapotranspiration
EWI	Equal Width Increment, a stream sampling method
FS	Fine sediment (< 16 or 20 microns) by mass
FSP	number of fine sediment particles per unit volume
GIS	Geographic Information System
GRMSE	Gilroy's estimate of Root Mean Square Error (see Glossary)
l	Liter
LRWQCB	Lahontan Regional Water Quality Control Board
LSPC	Loading Simulation Program in C++ (Lake Tahoe Watershed Model)
LTADS	Lake Tahoe Atmospheric Deposition Study
LTBMU	Lake Tahoe Basin Management Unit
LTIMP	Lake Tahoe Interagency Monitoring Program
MAPE	Median absolute percent error
MDQ	Mean daily discharge
MT	Metric Tonne
NDEP	Nevada Division of Environmental Protection
NH_4	Ammonium

NO_3-N	Nitrate-Nitrogen; the analytical method includes nitrite-nitrogen, which is usually negligible
NTU	Nephelometric Turbidity Units
n/y	Number of Particles per Year
POM	Particulate Matter
PON	Particulate Organic Nitrogen
PP	Particulate Phosphorus
PSW	The Pacific Southwest Research Station of the US Forest Service
PWS	Period-weighted sample method of calculating load
Q	Discharge in volume per unit time, i.e. cubic ft per second
Q-wtd	Flow weighted
RMSE	Root Mean Square Error (see Glossary)
RSE	Residual standard error
s.d.	Standard deviation
SNPLMA	Southern Nevada Public Lands Management Act
SRP	Soluble Reactive Phosphorus
SS	Suspended Sediment
SSC	Suspended Sediment Concentration, usually in mg/l
stn	Station; generally a USGS gaging station used for flow measurement and sampling
SWRCB	State Water Resources Control Board (California)
TDP	Total Dissolved Phosphorus
TERC	Tahoe Environmental Research Center
THP	Total Acid-Hydrolyzable-Phosphorus
TKN	Total Kjeldahl Nitrogen (all organic nitrogen plus NH4 +)
TKN + nitrate	Total DissolvedNitrogen
TMDL	Total Maximum Daily Load
TON	Total Organic Nitrogen
TP	Total Phosphorus
TRG	Tahoe Research Group, predecessor to TERC
TRPA	Tahoe Regional Planning Agency
TSS	Total Suspended Sediment; measured by a different method than SSC
UC Davis	University of California Davis
USACE	United States Army Corps of Engineers
USDA	United States Department of Agriculture
USEPA	United States Environmental Protection Agency
USFS	United States Forest Service
USGS	United States Geological Survey

Glossary

ACE An acronym for Alternating Conditional Expectations, which is the name of an algorithm that finds transformations of y and x that maximise the proportion of variation in y explained by x. *ACE* is implemented in package *acepack* of R (Spector et al., 2010). In our implementation of *ACE*, the algorithm was limited to finding monotonic transformations, in which y is an increasing function of x.

AIC Akaike's Information Criterion, specifically the corrected version, AIC_c recommended by Burnham and Anderson (2002). This criterion incorporates a penalty for including extra variables in the model.

AR Autoregressive model. A statistical model for time series data, where the response variable depends on its own previous value. In an AR model of order *p*, the response depends on its values in the previous *p* time steps. AR models may be used to describe regression residuals that are not independent but serially correlated.

AVAS An acronym for Additivity and VAriance Stabilization for regression, an algorithm that estimates transformations of x and y such that the regression of y on x is approximately linear with constant variance. AVAS is also implemented in package *acepack* of R, and we again limited the algorithm to finding monotonic transformations in which y is an increasing function of x. A more advanced version of AVAS is implemented in the *areg.boot* function in the *HMISC* package of R (Harrell, 2010). Transformed-variable models tend to inflate R^2 and it can be difficult to get confidence limits for each transformation. This method solves both of these problems using the bootstrap with monotonic transformations.

Bonferonni Correction A correction applied to the p value in a test for significance to take account of the expected number of false positives in a large number of tests. For example, if you test for a treatment effect on 100 plots, and set the p threshold at 0.05, then you can expect to find about 5 false positives. See Miller, 1981.

Categorical variable A variable that takes values from a fixed, usually small, number of specific outcomes or levels. As an example, substances can be classified as liquid, solid, or gaseous. In contrast, a continuous variable can assume an infinite number of possible values, e.g. all positive real numbers.

Duan's smearing correction for retransformation bias When predictions are made from a regression model for a transformed response such as the logarithm of concentration, they generally need to be retransformed back to the original units. A bias is introduced by the retransformation: the retransformed response is not equal to its mean value for the given set of predictors. There are three main methods for adjusting the retransformed responses to remove the bias (Cohn et al., 1989). Two of these methods assume that the transformed responses are normally distributed. The third method, Duan's smearing correction, does not. To apply Duan's smearing correction each prediction is multiplied by the mean of the retransformed regression residuals.

Hysteresis In the context of streamflow and water quality, hysteresis implies that the two are out of phase. Most commonly, concentration peaks before streamflow and a graph of concentration versus discharge forms a loop pattern, referred to as clockwise hysteresis, in which concentration is greater for a given discharge on the rising limb of a hydrograph than for the same discharge on the falling limb. Seasonal hysteresis can also occur, where concentration is greater for a given discharge early in the season than for the same discharge later in the season.

GRMSE Gilroy's (Gilroy et al., 1990) estimate of root mean square error (GRMSE) for the predicted load. This criterion utilizes information in the complete predictor data set that is used to compute a sediment load.

MK Test The Mann-Kendall trend test is a non-parametric test often used to test for the significance of a time trend in water quality and air quality data. It has some advantages over ordinary least-squares regression in that it does not assume normality of residuals, and is resistant to leveraging by outliers. The test calculates the slope of all lines between all possible pairs of points in a time series, and finds the median slope. See Helsel and Hirsch (2002).

LOESS An abbreviation for LOcal regrESSion, otherwise known as locally-weighted scatterplot smoothing (Cleveland and Devlin, 1988). A smooth curve is generated by fitting weighted first or second degree polynomials to neighborhoods around each data point. Each local regression gives more weight to points near the point whose response is being estimated and less weight to points further away. A user-specified "smoothing parameter" determines how much of the data is used to fit each local regression. The R implementation was employed using first degree polynomials, smoothing parameter equal to 0.8, and two different options for curve-fitting: (1) "gaussian" fitting by least-squares (loess.g), and (2) "symmetric" fitting (loess.s), which uses a "re-descending M estimator with Tukey's biweight function."

MAPE Median Absolute Percent Error. The 50th percentile error, as percent of the sample mean. This was used as a diagnostic criterion for comparing load calculation programs.

Monotonic transformation A variable transformation that preserves the ordering of values. For example, a logarithmic transformation is monotonic because $x_2 > x_1$ guarantees that $\log(x_2) > \log(x_1)$.

Monotonic trend A chronological series of values that never decreases (monotonically increasing series) or that never increases (a monotonically decreasing series).

Power of a test The power of a test is its ability to detect a treatment effect or trend, that is, to detect an alternative hypothesis when it is true.

Primary Stations. The stream gaging and sampling stations located near the lake, at or slightly up-stream from tributary mouths. Secondary stations are located at higher elevations, and were installed later with the goal of sampling relatively undeveloped areas.

PWS Period-weighted sample method of calculating total load. Each two successive concentrations are averaged, multiplied by the cumulative discharge between sampling times, and the resulting load increments summed over the water year.

Resampling Drawing independent sample set from the same population (either with or without replacement

RMSE Root Mean Square Error is a frequently used measure of accuracy, i.e. the differences between values predicted by a model or an estimator and the values actually observed. It is calculated as the square root of the mean of the squared differences between observed and predicted values.

RSE Residual standard error is the estimate of the standard error of the regression residuals

Abstract

Since 1980, the Lake Tahoe Interagency Monitoring Program (LTIMP) has measured discharge and sampled water quality at up to twenty stations in Tahoe basin streams. Measured constituents have included suspended sediment and various forms of nitrogen and phosphorus. In order to improve the usefulness of the program and the existing data base, we have 1) identified and, to the extent possible, corrected for two sources of bias in the data base; 2) generated synthetic data sets using turbidity, discharge, and time of year as explanatory variables for different forms of nitrogen, phosphorus and suspended and fine sediment; 3) resampled the synthetic data sets and part of the historic record in experiments to compare the accuracy of different load calculation models; 4) identified the best load calculation models for estimating each constituent load; 5) using the best models, recalculated total annual loads for all constituents and stations over the period of record; 6) regressed total loads against total annual and maximum daily discharge, and tested for time trends in the residuals; 7) developed a spreadsheet showing the relationships between sample size and confidence limits for estimated loads of the constituents. The time series analyses show significant long-term downward trends in some constituent loads and some streams, which we attribute to long-term recovery of watersheds from historic disturbance. The confidence-limit spreadsheet will be a useful tool for managers in modifying LTIMP to make it a streamlined and effective tool for future water quality monitoring in the Tahoe Basin.

Executive Summary

Since 1980, the Lake Tahoe Interagency Monitoring Program (LTIMP) has measured discharge and sampled water quality at up to twenty stations in Tahoe basin streams. Suspended sediment and various forms of nitrogen, phosphorus and iron have been measured. The purpose of the program has been to document long-term trends in water quality of the major tributaries to the Lake, and thus provide a basis for public policy and management decisions that may affect Lake clarity.

Several problems have limited the usefulness of LTIMP and over time, some of these problems have become more apparent. Based on our current understanding of water quality and its effect on lake clarity, these problems include 1) analysis of water quality parameters with a limited connection to lake clarity; 2) changes in the chemical methods and chemical species analyzed, which complicate efforts to measure long-term trends; 3) changes in the time of sampling for some streams, which may have introduced bias in records for flow-driven constituents such as suspended sediment and total phosphorus; 4) use of inaccurate models and lack of estimates of statistical error in calculating total constituent loads.

The purpose of this project is to review the existing program, develop and apply procedures for recalculating total constituent loads and time trends, and recommend programmatic changes that will make LTIMP a useful and cost-effective management tool. We take both a retrospective approach, looking back at the existing program and data, and a forward-looking approach, recommending changes based on calculated sample numbers and project costs for different levels of uncertainty in estimates of total constituent loads.

<u>The Problem of Bias</u>

Two sources of bias have crept into the LTIMP record. First, the nitrate-N record has been affected by changes in chemical methods. From 1976 to April 2003, nitrate-N was analyzed by reducing it to nitrite, and developing a color for photometric analysis. In 2003, it was discovered that addition of a catalyst--pyrophosphate with copper--gave higher yields, and this catalyst was used in subsequent analysis. In order to provide a basis for adjusting the old data, nitrate-N was measured by both the old and new methods in 2,370 pairs of samples from all LTIMP stations, between 2003 and 2008. A test for homogeneity of the regression coefficients showed significant differences between stations, possibly because the concentrations of interfering cations varied between stations. Separate regression equations for each station were thus used to adjust the old data to the value estimated for the new method. Details of the adjustment procedure are given in Appendix A-1.

The second source of bias is related to a change in the times of sampling throughout the day. Prior to 1989, USGS staff sampled intensively during snowmelt, with samples collected both day and night. Due to (very real) safety concerns, nighttime sampling was cut back or discontinued for streams in Nevada, as well as Trout Creek and the Upper Truckee River (UTR). The TRG/TERC, however, has continued nighttime sampling on Ward, Blackwood and General Creeks. On the larger watersheds, especially in the latter days of the snowmelt season, the daily snowmelt pulse (which carries most of the daily water volume and constituent loads) may arrive at the gaging station after dark. Eliminating nighttime sampling could thus create an apparent downward trend or sudden drop in total load or concentration. In small watersheds, daytime-only sampling could over-sample the rising hydrograph limb, and result in an upward bias in estimated daily load. Although there is no way to reconstruct the nighttime data that have been lost by the reduction in nighttime sampling, we were able to estimate the potential magnitude of bias that would have been introduced in Ward, Blackwood and General Creeks by calculating loads with and without nighttime samples, and for other streams by resampling simulated records. At Ward Creek, elimination of nighttime sampling would create an apparent downward trend or sudden drop in suspended sediment. The resampled simulated records showed modest introduced bias in some constituents for some stations. Because of these problems, time trends in total load estimates were calculated for daytime only samples as well as for all samples.

<u>Testing and Choosing Load Estimation Methods</u>

To develop a basis for selecting the best load calculation methods, we used a simulation approach. The simulations involve resampling from (1) synthetic populations of target variables or (2) worked records developed from stations and years (mid-1980s) when sample sizes were large. The synthetic data sets were constructed using regression relationships with transformed and untransformed continuous variables (discharge, lagged discharge, turbidity, conductivity, water temperature) from previous studies. Random error was added to the synthetic concentrations. For maximum relevance to historic load estimation, the data sets were resampled in a way that retains the characteristics of historic LTIMP sampling protocols, and loads were estimated using multiple methodologies on each synthesized sample. The worked records were resampled by subsampling days on which LTIMP water quality samples were collected.

To create the synthetic data sets, we obtained data sets from several researchers who have conducted studies in sub-basins around the lake: Trout Creek (Gary Conley, 2nd Nature LLC), Homewood Creek (Mark Grismer, 2013), Angora Creek (Alan Heyvaert), Ward and Blackwood Creeks (Andrew Stubblefield), Rosewood and Third Creeks (Rick Susfalk). An essential feature of all these data sets is that they included near-continuous turbidity data that could be correlated with sample data to produce near-continuous concentration data. In addition to the turbidity data, several stations included near-continuous conductivity or water temperature data. All these variables, together with discharge, gave us the ability to develop models for accurately, or at least realistically, constructing synthetic data sets that we could use for simulating sampling protocols and load estimation methods.

We then resampled the synthetic data sets 90 times each for 5 sample sizes between 10 and 80, and estimated the load using a variety of regression models, both parametric and non-parametric, and interpolating methods (Table ES-1). The variables used in the regression models included turbidity (T), instantaneous discharge (Q), mean daily discharge (MDQ), mean daily discharge lagged by 1 day (MDQ_1), and number of days since start of water year (D).

Table ES-1. Load estimation methods used in simulations. See Glossary for definitions.

Short name	Type	Time step	Description
rcload.turb	simple regression	30-min	log(C)~ g(T)
rcload.turb2	simple regression	30-min	$C^{0.5} \sim T^{0.5}$
rcload	simple regression	30-min	log(C) ~ log(Q)
rcload.mdq	simple regression	daily	log(C) ~ log(Q)
rcload.mdq2	multiple regression	daily	log(C) ~ log(Q) + log(MDQ/MDQ_1)
rcload.mdq3	multiple regression	daily	log(C) ~ log(Q) + log(MDQ_1)
rcload.mdq4	multiple regression	daily	log(C) ~ log(Q) + log(MDQ/MDQ_1)+D
rcload.mdq5	multiple regression	daily	log(C) ~ log(Q) + D
rcload.mdq6	simple regression	daily	$C^{0.5} \sim Q^{0.5}$
loess.g	nonparametric	daily	Loess with gaussian fitting
loess.s	nonparametric	daily	Loess with symmetric fitting
ace	nonparametric	daily	ACE transformations (see text)
avas	nonparametric	daily	AVAS transformations (see text)
areg.boot	nonparametric	daily	AVAS *areg.boot* implementation
pdmean	averaging	daily	Period-weighted sampling estimator
pdlinear	interpolating	daily	Daily interpolator
pdinstant	interpolating	30-min	"Continuous" interpolator
pdlocal2	interpolating	daily	Two-point rating curves + global curve
pdlocal2a	interpolating	daily	Two-point rating curves + interpolation
pdlocal4	interpolating	daily	Four-point rating curves + global curve
rcb1	best regression	daily	Selection by AIC without turbidity
rcb2	best regression	daily	Selection by GRMSE without turbidity
rcb3	best regression	daily	Selection by AIC, all log(C) models
rcb4	best regression	daily	Selection by GRMSE, all log(C) models

The best model was selected based on both root mean squared error and median absolute error, averaged over all synthetic data sets for each sample size. Note that the last four methods in Table ES-1 are themselves programs that select the best regression model among five models (rcload.mdq through rcload.mdq5), using one of two selection criteria. Table ES-2 shows the final selections of best methods. "Period Weighted Sampling Estimator" or "Period Weighted Mean" (pdmean) is an averaging method that steps through the record, averaging successive pairs of concentration values and multiplying the average by the total discharge between the two sampling times. It is sometimes called the "integration method." Selection by GRMSE (rcb2) selects the best regression of 5 models for the logarithm of concentration using Akaike's Information Criterion (AIC).

Table ES-2. Selected best estimation methods for all constituents with and without turbidity data.

Constituent	***With turbidity***	**Without turbidity**
SS	rcload.turb2	rcb2
FS	rcload.turb2	rcb2
FSP	rcload.turb2	rcb2
TP	rcb3	rcb2
TKN	pdmean	pdmean
NO3	pdmean	pdmean
SRP	pdmean	pdmean

<u>Recalculating Total Loads</u>

The selected methods (Table ES-2) were then used to recalculate annual loads for all stations, years, and constituents. Nitrate-N data were corrected for bias, as explained above. Loads were only calculated when a minimum of 10 samples was available for analysis. For each station, year, and water quality parameter, we saved the following information, which is provided with this report in electronic form as “LTIMP loads.xls”:

- Annual yield computed from all samples collected
- Annual flow volume
- Annual peak daily flow
- Sample size
- Discharge weighted mean concentration
- In the case of best regression, the form of the best model and Gilroy’s estimate of root mean squared error (GRMSE)
- Annual yield computed from daytime samples only (9am-6pm)
- Daytime sample size

<u>Time trends in Annual Loads</u>

Hydrology—especially annual runoff volume and the maximum daily flow—plays a major role in explaining the variance in total annual load. To examine time trends, we ran regressions of annual load vs. these two hydrologic variables, and included station as a categorical variable. We used the daytime samples only, to avoid the problem of trends induced by the time-of-sampling bias. We then tested separately for time trends in the regression residuals from each station, using the Mann-Kendall Trend Test (see Glossary). With 20 separate stations tested, we used the Bonferroni correction to limit the number of false positives, setting the threshold significance level at 0.0025 for each test. The significant trends identified were:

- SS: ED-3, IN-1, TH-1, UT-1
- TP: GC-1, IN-2, TH-1, UT-1, WC-8

- TKN: ED-3
- NO3-N: BC-1, GC-1, IN-1, IN-3, TH-1, UT-1, WC-8
- SRP: GC-1, TC-1, TH-1

All significant trends are downwards, with the exception of SRP at TC-1. The significant trends at IN-2 and ED-3 are for short periods of record. Those sites have not been sampled since 2006 and 2001, respectively. Figure ES-1 shows the nitrate-N data for all stations. Similar figures for other constituents are in the main body of this report.

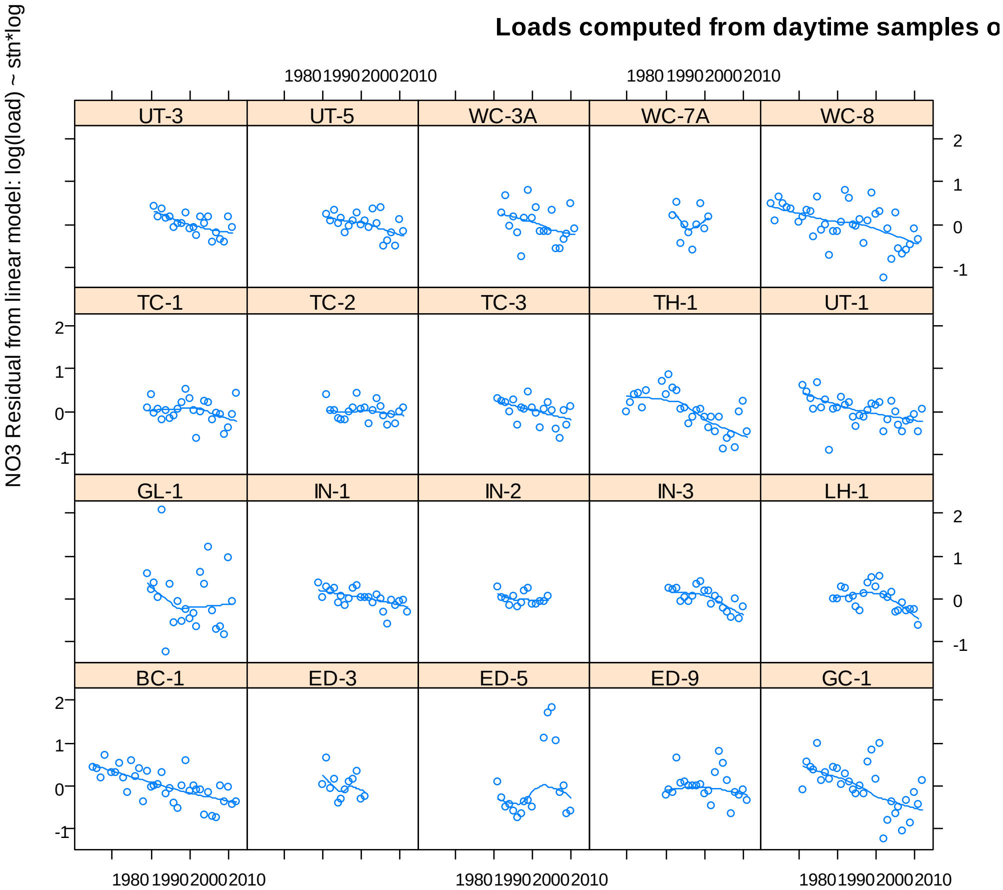


Figure ES-1. Trends in NO_3-N after accounting for inter-annual variation in total and maximum daily runoff

The occurrence of so many downward trends in loads, especially for NO_3-N, is striking. We hypothesize that the trends are caused by long-term recovery from logging and overgrazing in the 19th century and first half of the 20th century. Essentially, the forests are accumulating biomass, and becoming more effective in retaining nitrogen, phosphorus and sediment. In the case of Blackwood Canyon, recovery may involve a shift from nitrogen-fixing alder toward conifers, which produce a litter and humus layer with high carbon-nitrogen ratio (Coats et al., 1976). The long-term trend toward warmer temperatures could accelerate plant growth and contribute to closing of nutrient cycles and reductions in sediment production. Land use controls implemented since the mid-1970s may also have played a role.

Confidence Limits and Costs for New Estimates of Total Load

The simulations of concentration and calculations of total load allowed us to examine the relationships between sample size and error in new load estimates for each constituent. We created a spread-sheet showing alternative levels of precision for each constituent as a function of sample size (Table 13-1). The sample numbers refer to the required number of samples per year at a given station. For example, with 25 samples per year, one can be 90 percent sure that the true annual load of total phosphorus is within +/- 20 percent of the value estimated using the best model selected by the GRMSE criterion. For the same level of confidence and percent error for SSC, 67 samples per year would be required. The relationship between confidence level and sample number is highly non-linear. To achieve the 90/10 level for TP and SSC would require over 100 samples per year.

Table 13-1 also shows the comparison of required sample numbers for an improved LTIMP (using new load calculation models) with a turbidity-based program, where loads of the particulate constituents are calculated by regression with continuous turbidity. For SSC at the 90/20 level, the required sample size drops from 67 to 20.

These sample size tables have been used to develop cost estimates for different levels of effort in a realigned LTIMP. The spreadsheet showing the cost alternatives is available on request.

The derived confidence limits can be used in the design of monitoring projects to calculate the sample size required to detect a given percent change in load between an upstream and downstream station. For example, without turbidity data, detecting a 30% change in total phosphorus load would require at least 40 samples along with instantaneous and mean daily discharge.

Summary and Recommendations

In this study, we have developed and compared different methods of calculating total constituent loads, and expressed the results as the number of samples (per station-year) required to achieve a given level of confidence that the true load is within a given error band around the estimated load. Using the best methods (that is, the methods that maximize precision and minimize bias), we recalculated the total annual loads of NO_3-N, NH_4-N, TKN, SRP, TP and SSC for all of the LTIMP stations over the periods of record. We then related the annual loads to annual runoff and maximum daily peak discharge, and (for all but NH_4-N) analyzed time trends in the residuals. The significant downward trends indicate some long-term improvement in water quality, which we suggest may be due to long-term recovery of terrestrial ecosystems from 19th and 20th century disturbance.

Based on our results and our experience working with the LTIMP data, we recommend the following:

- Near-continuous measurement of turbidity and temperature with automated probes should become a central part of the realigned LTIMP.

- Fine sediment is now recognized as an important factor in lake clarity and is incorporated in the TMDL targets and used in the Lake Clarity Model. Increased emphasis should be placed on its measurement as number of particles.

- The load calculation models developed in this study present an opportunity for major cost savings (or improvements in accuracy of load estimates) especially for the dissolved constituents.

- If station numbers need to be reduced for budget reasons, stations on the big contributing streams (Ward, Blackwood, Trout Creeks and the UTR) should have priority for continued discharge measurement and sampling.

- An intermediate confidence and error level—90 percent confidence that the true value lies within +/- 20% of the estimated value—is achievable with 20 samples per year combined with continuous turbidity, for all constituents, and 25 samples per year without turbidity, for the chemical constituents. Without good turbidity data, this level for SSC would require about 70 samples per year.

- The relationships between sample size, confidence and error should be used to plan monitoring of restoration and mitigation projects.

- Ammonium-N could be dropped from the list of constituents measured, since about half the time its concentration is below the MDL. It is included in the measurement of TKN.

- The LTIMP needs a real director, with a strong background in hydrology, biogeochemistry and statistics along with experience in fund-raising. The director should be housed in one of the scientific research organizations active in the basin (TERC, USGS, DRI/UNR or PSW), and have decision-making authority on operational matters. Overall policy direction would continue to be the responsibility of the management agencies. With continued involvement of support staff from the scientific organizations, this could perhaps be a half-time job. If responsibilities include the urban runoff, air quality and lake monitoring programs, it would be at least a full-time job.

1 Introduction and Background

The Lake Tahoe Interagency Monitoring Program (LTIMP) was established in 1979 in response to declining water quality and clarity of Lake Tahoe. Its original purpose was to provide data for estimating annual streamflow and loads of nutrients and sediments from basin watersheds to the Lake. It is currently funded and administered by the US Geological Survey (USGS), the University of California Davis Tahoe Environmental Research Center (TERC), the Tahoe Regional Planning Agency and the US Forest Service Lake Tahoe Basin Management Unit (LTBMU). These agencies, along with the Lahontan Regional Water Quality Control Board (LRWQCB, or "the Board"), have statutory responsibility for maintaining and improving water quality in the Tahoe basin. To fulfill their responsibilities, they need accurate and precise information on the loading rates and sources of sediment and nutrients that enter the Lake, along with assessment of the status and long-term trends in basin water quality. In addition, the data provided by LTIMP are of considerable interest to scientists, and to engineers, planners and consultants in the private sector. The LTIMP data base has proven to be especially important in in the development of the TMDL (Total Maximum Daily Load) Program (Lahontan and NDEP, 2010), the assessment of climate change impacts on the Lake (Sahoo et al., 2013) and in a number of scientific journal articles.

Funding levels for LTIMP have varied over the years, from a maximum of ______ in ___to _______ in FY 2013. Since 1988, the USGS and TRPA have spent over $9.1 million on the program (Alvarez, et al, 2007. Given the present fiscal environment, however, it is likely that future funding levels will not be generous. In order to maintain a program that will provide the information needed by managers and scientists in the basin, it will be important to establish the relationships between costs and the quality of data, including the accuracy and precision of loading estimates, and to choose the most cost-effective methods for data acquisition.

This purpose of this project is to develop and analyze alternatives for realigning LTIMP to provide useful data for the agencies and research scientists. We first take a retrospective approach. This includes 1) cleaning up and removing (where possible) identifiable sources of bias in the data; 2) using a Monte Carlo method to find the optimum methods for calculating total loads; 3) recalculating loads for all years, constituents and stations; 4) using the recalculated loads to test for time trends in water quality. Second, we evaluate alternative sampling protocols for both cost and confidence limits of loading estimates, in order to inform future decisions about funding and project design. Our approach is to outline the costs and benefits of various levels of effort for the future program, recognizing that the agencies themselves must choose how much to spend and what level of accuracy is acceptable. We do, however, make some recommendations of our own.

The early phase of the LTIMP was an outgrowth of stream sampling on Ward and Blackwood Creeks by the UC Davis Tahoe Research Group (TRG) and USGS (Leonard et al., 1979. Early data from that effort have been incorporated into the LTIMP data base, although instantaneous discharge values are not available. In 1978 the sampling was funded, formalized, and extended to four more streams on the California side, with USGS gaging stations installed at the stream mouths near the lake. In 1989~~,~~ upper watershed stations were added in five tributaries to provide (in some cases) better data for catchment areas not directly affected by development. Table 1-1 lists the stations that have been included in the LTIMP, and indicates span of years of sampling. Figure 1-1 shows the locations of the stations that are listed in Table1.1. Note the significant cutback in the number of stations starting in 2011.

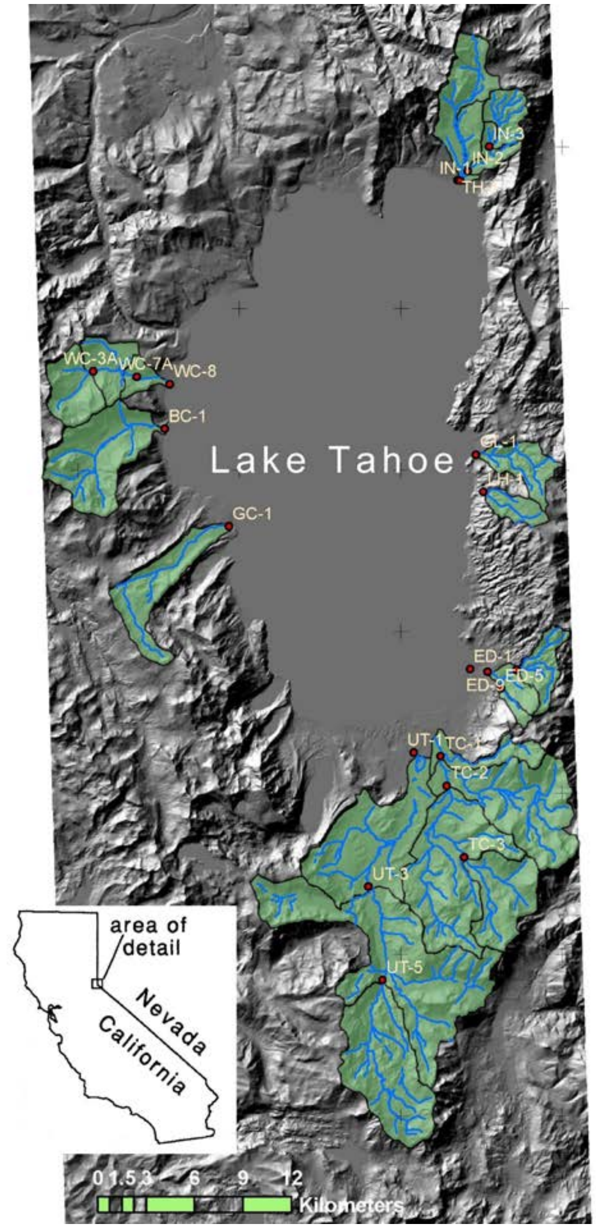


Figure 1-1. Map of the Tahoe Basin, showing locations of the LTMP watersheds and sampling sites.

Table 1-1. List of LTIMP stations with USGS code, LTIMP code and range of years sampled. A few stations are missing some years, not shown.

			Water Year		
Tributary Name	**Sta. Name**	**USGS Sta ID No.**	**Begin Record**	**End Record**	**Length Record**
Angora Cr.	AC-1	1033660958	2007	2008	2
Angora Cr.	AC-2	103366097	2008	2010	3
Blackwood Cr.	BC-1	10336660	1974	2012	39
Edgewood Cr.	ED-1	10336765	1984	2002	19
Edgewood Cr.	ED-2	10336761	1984	1992	9
Edgewood Cr.	ED-3	103367585	1989	2002	14
Edgewood Cr.	ED-4	10336750	1989	2002	14
Edgewood Cr.	ED-5	103367592	1990	2011	22
Edgewood Cr.	ED-6	10336756	1981	2001	21
Edgewood Cr.	ED-9	10336760	1992	2011	20
First Cr.	FI-1	10336688	1970	2002	33
General Cr.	GC-1	10336645	1980	2012	33
Glenbrook Cr.	GL-1	10336730	1972	2011	40
Glenbrook Cr.	GL-2	10336725	1989	2000	12
Incline Cr.	IN-1	10336700	1970	2012	43
Incline Cr.	IN-2	103366995	1989	2006	18
Incline Cr.	IN-3	103366993	1989	2011	23
Incline Cr.	IN-4	12336700	1989	1990	2
Incline Cr.	IN-5	103366997	1989	2002	14
Loganhouse Cr.	LH-1	10336740	1984	2011	28
Loganhouse Cr.	LH-3	10336735	1991	2002	12
Rosewood Cr.	RC-2	103366974	2001	2003	3
Second Cr.	SE-1	10336691	1991	2001	11
Second Cr.	SE-2	103366905	1995	2000	6
Snow Cr.	SN-1	10336689	1980	1985	6
Taylor Cr.	TA-1	10336628	1998	1999	2
Trout Cr.	TC-1	10336790	1972	2012	41
Trout Cr	TC-2	10336775	1989	2011	23
Trout Cr	TC-3	10336770	1990	2011	22
Trout Cr	TC-4	10336780	1974	2002	29
Third Cr.	TH-1	10336698	1970	2012	43
Third Cr.	TH-4	103366965	1989	2000	12
Third Cr.	TH-5	103366958	1989	2001	13

Upper Truckee	UT-1	10336610	1970	2012	43
Upper Truckee	UT-2	103366098	1989	2002	14
Upper Truckee	UT-3	103366092	1989	2011	23
Upper Truckee	UT-4	103366094	1989	2002	14
Upper Truckee	UT-5	10336580	1989	2011	23
Ward Cr.	WC-3A	10336674	1991	2011	21
Ward Cr.	WC-7A	10336675	1989	2003	15
Ward Cr.	WC-8	10336676	1972	2012	41
Wood Cr.	WO-1	10336694	1970	2002	33
Wood Cr.	WO-2	10336692	1991	2001	11

The water quality constituents measured in streamwater samples have focused on suspended sediment and different forms of nitrogen, phosphorus and iron. Table 1.2 lists the specific forms analyzed. For details on the analytic methods used, see Liston et al., 2013. Note that different methods have been used for nitrate-N, total phosphorus and iron. These changes in chemical methods may have introduced bias that could affect apparent time trends in loads and concentrations. The problem of bias introduced by a change in the method for nitrate-N is discussed in Appendix A-1.

Table 1-2. Water quality constituents analyzed at some time during the LTIMP

ASSAY CODE	ASSAY DESCRIPTION
DBAFE/DFE	Iron, biologically available, diss. (ug/l as Fe)
DHP	Phosphorus, hydrolyzable+ortho, diss. (ug/l as P)
DKN	Nitrogen, ammonia plus organic, diss. (ug/l as N)
NH_4-N	Nitrogen, ammonia, dissolved (ug/l as N)
NO_3-N	Nitrogen, NO2 + NO3, dissolved (ug/l as N)
DP	Phosphorus, dissolved (ug/l as P)
DRFE	Iron, dissolved (ug/l as Fe)
Sol.Fe	Soluble Iron (ug/l)
SRP	Phosphorus, orthophosphate, diss. (ug/l as P)
TBAFE/BAFE	Iron, biologically available, total (ug/l as Fe)
THP	Phosphorus, hydrolyzable+ortho, total (ug/l as P)
TKN	Nitrogen, ammonia + organic, total (ug/l as N)
TNH_4-N	Nitrogen, ammonia, total (ug/l as N)
TNO_3-N	Nitrogen, NO2 + NO3, total (ug/l as N)
TP	Phosphorus, total (ug/l as P)
TRFE	Iron, total (ug/l as Fe)
TRP	Phosphorus, orthophosphate, total (ug/l as P)
SSC	Suspended Sediment Concentration, mg/l
TSS	Total Suspended Sediment, mg/l

Changes in the regulatory environment have imposed changes in demands for information on the LTIMP. Under Sec. 303(d) of the Clean Water Act, the LRWQCB found eight basin streams out of compliance with the standards of the Basin Plan, and the Nevada Division of Environmental Protection added three more. Table 1-3 lists the streams and the constituents for which the streams are listed.

In 1995, the LRWQCB in its revised Basin Plan defined the water quality of objective for Tahoe basin streams such that the 90^{th} percentile value (exceeded by 10 percent of the samples) should not exceed 60 mg/l. The 90^{th} percentile cannot be defined without a set of samples, and what the samples should represent is not defined. For Blackwood Creek, using all of the available samples in the data base, the 90^{th} percentile concentration is 169 mg/l. But the sampling times were concentrated during high flows, when sediment loads were highest. If the 90^{th} percentile is based on estimated values for all days in the record (available for 1976-87), the 90^{th} percentile concentration is 20 mg/l. A change in the sampling regime for sediment might require a revision of the Basin Plan objective.

In 2010, the Board formally adopted a TMDL, which identified four major source types for pollutant loads, and set reduction targets aimed ultimately at improving lake clarity. The source types include urban runoff, channel erosion, atmospheric deposition, and upland forested areas. The LTIMP stations, due to their locations, sample the contributions primarily from upland forested areas and stream channels, with contribution in some watersheds from urban sources. The contribution from the latter is targeted more directly by a separate monitoring program using automated sampling methods.

Table 1-3. 303(d) listed streams, and parameters named as impairing water quality.
Lists from LRWQCB (1995) and NDEP (2013)

	Parameter					
	Total N	Total P	Sediment	Total Fe	Zn*	Pathogens
Ward	X	X	X	X		
Blackwood	X	X	X	X		
General		X		X		
UTR						
Trout	X	X		X		X
Edgewood		X		X		
Glenbrook		X				
Incline				X		
First					X	
Second					X	
Third					X	

*May be due to contamination; reevaluation pending.

Since its inception in 1978, the administration and funding sources of LTIMP have varied. In its first five years eleven state and federal agencies contributed funds or made in-kind contributions of labor or facilities. During that period, the State Water Resources Control Board (SWRCB) provided the chairmanship, but actual implementation was the responsibility of the UCD Tahoe Research Group. In 1988, the Carson City office of USGS Water Resources Division took over the federal role from the Sacramento office and data reporting was incorporated into the USGS/STORET system. In 1992 additional stations were installed in upper watershed locations.

2 Current Understanding of Water Quality Issues in the Tahoe Basin

Our current understanding of basin water quality issues and their relationship to the clarity of Lake Tahoe is somewhat different from our understanding at the time the LTIMP was established. This is due in part to the monitoring program itself and in part to related research projects over the last 34 years. Realignment and redesign of the LTIMP must be informed by our current understanding of the water quality issues. Here we highlight the important issues that relate to the design of the monitoring program. For a deeper look into these issues, the reader is referred to Rowe et al., 2002.

2.1 Limiting Nutrients

Nutrient addition experiments during the 1960s indicated that primary productivity in the lake was limited by nitrogen. Additions of phosphorus had little stimulatory effect (Goldman and Armstrong, 1969; Goldman and De Amezaga, 1975). By 1980, however, phosphorus limitation was noted, with some additional role played by iron and zinc (Goldman, 1981). Subsequent work has confirmed the increasing importance of phosphorus as a limiting nutrient, with a shift from co-limitation by N and P to consistent P limitation (Goldman et al, 1993). Chang et al. (1992) also found a strong limitation of primary productivity by phosphorus as well as a possible role for iron. They noted, however, that the availability of the latter is strongly influenced by complexation with dissolved organic matter.

The primary cause of the shift from N to P limitation has been the high rate of nitrogen loading to the lake. The most important source of nitrogen is atmospheric deposition of nitrate and ammonium. Jassby et al. (1994) found that the combined wet and dry deposition of dissolved inorganic nitrogen (DIN) on the lake (1989-1991) was 19 times that of the loading from the watersheds. This finding contributed to increased attention to the role of air pollution in modifying water quality.

Since WY 1989, the LTIMP has monitored concentrations and fluxes of dissolved and particulate organic nitrogen from the watersheds. The results show that organic nitrogen (measured as Total Kjeldahl Nitrogen) typically accounts for about 85 percent of the total nitrogen delivered to the lake from the watersheds (Coats and Goldman, 2001; Coats et al., 2008). Between 40 and 80 percent of the Total Organic Nitrogen (TON) load is dissolved (DON). If only a fraction of the DON is biologically available, its importance as a nitrogen source for algae growth would outweigh that of dissolved inorganic nitrogen (DIN). In a study of the bioavailability of DON to bacteria and phytoplankton, Seitzinger et al. (2002) found that 30-45 percent of the DON in streamwater from a pine forest was biologically available. The availability of DON from urban runoff varied from 48 to 70 percent. The bioavailability of even humic-associated nitrogen in a river draining coniferous forest (generally thought to be refractory) may be as high as 37 percent (Carlsson et al. 1999). Although primary productivity in the lake is now P-limited, the warming of the lake may (by the end of this century) increase the internal supply of phosphorus, and thus shift the lake back to a condition of N-limitation (see Sec. 2.5, below). Although the bioavailability of DON has not yet been measured in the Lake Tahoe or basin streams, TKN and DON are clearly important parts of the lake's nutrient budget.

2.2 The Importance of Fine Sediment

At its inception, LTIMP was focused primarily on nutrients known to be limiting to primary productivity in the lake. In recent years, however, modeling, field and laboratory work have shown the importance of fine (0.5-16 µm) sediment (FS). According to Swift, et al. (2006), about 60 percent of the loss of clarity (measured by Secchi depth) is caused by fine sediment. Four streams—Ward, Blackwood, Trout Creeks, and the UTR account for about half of the stream contributions to the Lake (Nover, 2012). Fine sediment is doubly important, in that it not only reduces lake clarity directly, but also carries adsorbed phosphorus that may be liberated to support algal growth (Froelich, 1988; Hatch et al., 1999). Methods for sampling, measuring and reporting fine sediment loads and concentrations are discussed below in Section 4.1.

2.3 Channel Erosion as a Source of Sediment

The Tahoe TMDL (Lahontan and NDEP, 2010; Simon et al., 2003; Simon, 2006) modeled the contribution of stream channel erosion to the lake's sediment load. They found that channel erosion accounted for 17 percent of fine sediment particles (< 16 µm). The models assumed, however, that the flood frequency distribution is stationary, that is, historic flood frequency can be used to estimate future flood frequency. With climate change such an assumption may be unreliable.

2.4 Land Use Influences: Urban Runoff, Fire and Dirt Roads

The Lake Tahoe TMDL (Lahontan and NDEP, 2010; Coats et al., 2008; Sahoo et al., 2013) showed that the nutrient and fine sediment contribution per unit area of urbanized lands far outweighs that of forested lands. Although urbanized areas account for only 6 percent of the surface runoff to the lake, they contribute 18, 7 and 67 percent respectively of the TP and TN loads and FS numbers. These estimates, reported by Sahoo et al. (2013) are based on extensive automated sampling and modeling of urbanized areas, sampling of atmospheric deposition, modeling of channel erosion, and previous LTIMP load estimates. Previous studies (Coats et al., 2008) identified unimproved dirt roads as a contributing factor in sediment production. This is consistent with a large body of work on impacts of timber harvest activities on sediment yields.

Two major fires in the last eleven years have provided information on the potential water quality impacts of wildland fire. The 600 ac Gondola fire of 2002 burned on steep slopes southeast of the lake. In order to document the water quality effects the USGS reactivated a retired stream gaging station on Eagle Rock Creek in the Edgewood Creek drainage. Sampling in the post-fire period showed large spikes in the concentrations and loads of suspended sediment, nitrate-N, ammonium-N, TKN, soluble reactive phosphorus (SRP) and TP (Allender, 2007). The results of stream sampling were corroborated by UNR study based on plot sampling in the burned and unburned area (Miller et al. 2007).

In 2007, the 3,100 ac Angora fire burned in mixed conifers, lodgepole pine and residential homes southwest of the lake. Monitoring of water quality was initiated shortly after the fire. Maximum concentrations of nitrate-N, TKN, SRP and SS were elevated relative to baseline concentrations for at least three years, but a downstream large wet meadow area trapped much of the pollutant load, and the impacts to the UTR and lake were slight (Reuter et al., 2012). The contrast in impacts between Gondola and Angora fires illustrates the site-specific nature of fire impacts on water quality, as well as the value of rapidly deployed post-fire monitoring.

2.5 Implications of Climate Change

The reality of climate change and its likely impacts in the Tahoe basin are by now well established. Documented historic and modeled future impacts in the basin include 1) a shift from snow to rain; 2) earlier spring onset of snowmelt; 3) increased frequency of large channel-modifying floods; 4) reduced summer low-flow; 5) more frequent and severe forest fires and tree mortality from bark beetle attacks; 6) warming of the lake, with concomitant increased thermal stability, food web changes and exotic species invasions; and 7) increased summer drought, especially on the east side of the basin (Coats, 2010; Coats et al., 2013; Sahoo et al., 2013, Riverson et al., 2013; Das et al., 2013; Whittmann et al., 2013).

These impacts of climate change have important implications for lake clarity and for future monitoring efforts in the basin. Increased thermal stability increases the resistance of the lake to deep mixing, which is essential for keeping the deep water of the lake oxygenated. The combined results from the LSPC, a basin-wide hydrology model (Riverson et al., 2013) and the lake clarity model (Sahoo et al., 2013) indicate that by the end of the century the deep mixing of the lake will cease, the bottom waters will become anoxic, and phosphorus that is locked up with oxidized iron (Fe^{+3}) in the lake sediments will be released to the water column. Ammonium would also be released from the sediment. The large release of phosphorus could trigger unprecedented algae blooms, and shift the trophic status of the lake from phosphorus limitation back to nitrogen limitation. This possibility underscores the importance of reducing phosphorus loading to the lake now, since much of the phosphorus entering the lake will be stored in the sediment for future release to the water column when deep-water sediments become anoxic.

Increases in frequency of large floods, intense wildfire and bark beetle mortality may increase surface and stream-channel erosion, contributing to the input of fine sediment to the lake. It will be important to monitor these processes and their contribution to sediment loads in order to design effective mitigation strategies.

Climate change may also have important effects on watershed biogeochemical processes. Warmer summer temperatures and higher CO2 may increase plant growth and the uptake of nitrogen and phosphorus, thus reducing the loss of these nutrients to the lake. Increased drought may reduce saturation in wetland soils and riparian zones, and thus reduce the denitrification of nitrate, and increase the immobilization of phosphorus with oxidized iron.

3 Measuring Stream Discharge

Measurement of stream discharge is an essential part of LTIMP. First, there can be no measurement of nutrient and sediment loads without discharge data. Second, discharge itself is important for water resource planning (water supply) and flood frequency analysis (flood control and land use planning).

Discharge measurement at USGS gaging stations in the Tahoe basin involves several steps. First, stage (or water surface elevation) is measured at 15-min intervals using a Sutron Accububble pressure transducer (http://www.sutron.com/products/Accububblegauge.htm) or similar instruments from WaterLOG (http://www.waterlog.com/index.php) . These devices emit a stream of air bubbles into the water, and measure the back-pressure with a pressure transducer. Data are stored in a data logger in the gaging station. The average hourly stage is transmitted to the GOES satellite, and is publically-available on-line (see for example http://waterdata.usgs.gov/ca/nwis/uv?site_no=10336645). Second, a rating curve is developed by measuring water velocity at selected points in a surveyed cross section, over a range of discharges. By multiplying cross sectional area by water velocity the relationship between stage and discharge is established. This curve, together with the stage record, is used to calculate both instantaneous and mean daily discharges. For calculation of instantaneous loads in the LTIMP, the instantaneous discharge data from the hydrologist must be collated with the concentration data from the laboratory. This is easiest if the date-time stamps in both records correspond exactly.

Problems in the discharge measurements sometimes arise, however, due to both short and long-term changes in the stage-discharge relationship, or problems in stage measurements. Short-term problems may occur due to ice forming at or just downstream from the cross section and creating a backwater effect, or interfering directly with the stage measurement mechanism by blocking the bubble outlet or increasing the water pressure above the outlet. Such problems can greatly increase the stage and apparent discharge. The ice effect can usually be identified by inspection of the gage record, though experience and skill are needed to remove the effect from the data. For this reason, stream gaging cannot be entirely automated. Long-term changes can occur if the surveyed cross section is scoured or filled at high discharge. Following a major flood, cross sections must be resurveyed in order to maintain a correct rating curve. For a more detailed description of stream-gaging methods, see: Turnipseed & Sauer, (2010) and Sauer & Turnipseed, (2010).

4 Water Quality Constituents: Sampling and Analytic Methods

Since its inception the LTIMP has made a number of changes in sampling methods, the constituents sampled, and the analytical methods used to measure them. Such changes are necessary in order to take advantage of changes in technology and in our understanding of the processes that influence lake clarity. The changes, however, create problems for analyzing long-term trends in water quality since detection limits, precision and accuracy may change with a change in methods. The following summarizes the constituents that have been included in the program, and discusses the rationale for including them.

4.1 Suspended Sediment

Suspended sediment samples are collected using the “equal-width increment” method at cross sections near the USGS gaging stations (Edwards and Glysson, 1999). The preferred method used by the USGS is to measure Suspended Sediment Concentration (SSC). In this method, samples are filtered onto tared filter paper, and the dried sample is weighed to determine mass of sediment from sample of known volume. In some cases (e.g. the Angora Creek study) the Total Suspended Sediment method has been used. In this method a sub-sample is drawn from the sample bottle by pipette and the concentration determined gravimetrically. This method works well when most of the sediment is silt or clay, but sand does not stay in suspension long enough to be adequately sampled by the pipette.

Fine (FS) sediment is not yet routinely measured in the LTIMP, though it has received more attention since its role in reducing lake clarity was recognized. The “fine” fraction is generally defined as particles < 16 or 20 µm, though the term is also applied to the fraction < 63 µm. Concentrations are sometimes reported as mass per unit volume, but reporting particle number per unit volume is preferable, since that variable is more closely related to light extinction and lake clarity.

At least three instruments are in use in the Tahoe basin for measuring the volumetric percentage of fine sediment or counting individual particles. The Micromeritics DigiSizer, the Beckman Coulter ls-13320, and the Particle Measuring Systems LiQuilaz instrument. The latter is the only particle counter of the three, it is the most sensitive, and is best suited for very low concentrations, such as those typical of Lake Tahoe water. The other two are best suited for stream and stormwater samples. Because of its sensitivity, the LiQuilaz is considered not suitable for measuring fjne sediment particle numbers at the high concentrations typical of stormwater runoff. Samples may require dilution, which can introduce an additional source of error (Heyvaert, et al., 2011; Nover, 2012; Rabidoux, 2005). Holding time of samples is an important issue, since flocculation after sampling may reduce the particle concentrations. Samples are sometimes "sonicated" with high-frequency sound to reduce flocculation (Nover 2012). Additional research on methods for measuring fine sediment and counting fine sediment particles in basin streams is urgently needed.

4.2 Turbidity

Turbidity is closely related to the concentration of fine sediment particles. It is not routinely measured in the LTIMP, but since it can be measured at short intervals (5-10 min) by a probe installed in the stream, it offers the possibility of developing virtually continuous records of some water quality constituents by regression. (Stubblefield et al., 2007; Heyvaert et al., 2011; Reuter et al., 2012). This approach shows most promise with constituents (FS, SSC, TP) that are physically related to turbidity. The available probes rely either on backscatter of light by suspended particles or on absorbance of light over a short path through the water column.

Successful use of turbidity probes, however, requires careful installation and frequent field maintenance (at 1-2 week intervals). Modern probes are equipped with small wipers to keep the light sensor free of biofouling by algae and bacterial slime, but these are not always effective. The probe must be installed so that sunlight will not cause erroneous readings (Lewis and Eads, 2009). Even with frequent maintenance, the probe data must be carefully inspected for problems. The program TTS Adjuster (available on disk as part of this report) allows the user to inspect turbidity data along with discharge data, and provides a useful tool for finding and correcting problems. As a backup, turbidity should always be read in the laboratory sample collected for chemical analysis, and compared with the simultaneous field value.

4.3 Nitrogen

Three forms of nitrogen are measured in the LTIMP: nitrate+nitrite-N, ammonium-N and Total Kjeldahl N. The latter includes ammonium-N; reported as Total Organic Nitrogen (TON), the ammonium fraction has been subtracted from TKN. Dissolved Inorganic Nitrogen (DIN) is mostly nitrate since ammonium-N is rapidly oxidized to nitrite and then to nitrate by nitrifying bacteria. Nitrate is measured by reducing it to nitrite, so values reported as nitrate are actually nitrate+nitrite.

4.3.1 Nitrate-nitrogen

Nitrate-N (NO_3-N) is readily available for plant uptake, and is relatively easy to measure. Sources in streamwater include atmospheric deposition, urban runoff, and decomposition of organic matter, especially from alder, which is an important nitrogen-fixer. Nitrate-N (or more correctly, nitrate + nitrite-N has been measured routinely in Tahoe basin streams since 1972, but the analytic methods have changed twice over the years. From WY 1972 through WY 1976, it was measured by reducing the nitrate to nitrite in columns packed with cadmium, and developing a color the intensity of which was read photometrically. In the fall of 1976, reduction by hydrazine replaced the use of the cadmium columns. It was later discovered that in streamwater (but not lake water) samples, divalent cations (Ca and Mg) interfered in the determination, and values were being under-reported. The problem was corrected by use of a pyrophosphate catalyst, and samples from each station were run with and without the catalyst to allow the earlier data to be corrected. Appendix A-1 describes the problem and shows the equations used to correct the old data. The corrected data were (based on data available at the time) used in the TMDL project (Lahontan and NDEQ, 2010; Coats et al., 2008) and more recent data are used in this project. The confidence limits on estimates of the individual corrected values are relatively wide, but the estimates should be unbiased and can be used to calculate total load.

Recent improvements in an optical probe for nitrate concentration offer the possibility of virtually continuous measurement of nitrate-N in basin streams (Pellerin, et al, 2013). Such a probe was used in a study of patterns of nitrate and dissolved organic matter in a stream in the Sleepers River watershed of Vermont (Pellerin et al., 2012). The minimum detection limit (MDL) is about 5 µg/l, only a little higher than the MDL of the currently-used nitrate method. By increasing the path of length of the light beam, an even lower MDL might be achievable. The probe, however, may be cost-prohibitive.

4.3.2 Ammonium-N

Ammonium-N (NH_4-N) has also been routinely measured since 1972, but about half the time the concentration is at or below the minimum detection level of about 2 µg/l. The main exception is Edgewood Creek, where concentrations have exceeded 200 µg/l in some years. Ammonium-N is routinely measured with TKN so it is included in the TMDL variable “Total Nitrogen”. To reduce costs it could be dropped from the list of constituents sampled without much loss of information.

4.3.3 Total Kjeldahl Nitrogen

TKN is measured by a micro-Kjeldahl method. Samples are digested in concentrated sulfuric acid, neutralized with NaOH, and the resulting ammonia distilled and captured for photometric analysis. Nitrate present in the original sample is not measured, but ammonium is. The MDL is about 2 ug/l. Most of the dissolved nitrogen in both the lake and streams is organic. From this relatively large pool, inorganic N is slowly mineralized and made available for plant growth. The rate of mineralization in basin streams and the lake is not known, but measurement of TKN will continue to be important in understanding the biogeochemical cycle of nitrogen in the lake.

4.4 Phosphorus

Phosphorus has been measured since the earliest days of the LTIMP and since it is the primary limiting nutrient for primary productivity in the lake its continued measurement will be important. Several forms have been measured and reported over the years; these include dissolved hydrolyzable phosphorus (DHP), dissolved phosphorus (DP), Total Hydrolyzable Phosphorus (THP), Soluble Reactive Phosphorus (SRP), Total Reactive Phosphorus (THP) and Total Phosphorus (TP). The measured concentration of these forms reflects the choice of filter size and method of digestion (if any). The different forms vary in their availability to algae. Of the various forms, TP and SRP have the longest and most complete records.

4.4.1 Soluble Reactive Phosphorus

SRP is virtually all orthophosphate (PO_4). It is readily available to algae and bacteria, and in the stream environment, its concentration is strongly controlled by chemical equilibrium reactions with iron and aluminum hydroxyoxides and calcium minerals (Froelich, 1988). For this reason, its concentration varies within a fairly narrow band, and is not closely tied to stream discharge.

Hatch et al., (1999) found that dissolved inorganic P (SRP) best represents the short-term stream bioavailable P.

4.4.2 Total Phosphorus

TP includes both soluble and insoluble forms. The latter include minerals weathered from bedrock such as calcium apatite and both dissolved and particulate organic matter. The rate at which P is released from mineral particles and organic matter is highly variable, but any form of TP may eventually be released and become available.

In some cases,Total Hydrolyzable P (THP) rather than TP has been measured. THP involved digestion of samples with sulfuric acid and spectrophotometric analysis of the resulting orthophosphate. The TP digestion uses acid persulfate, and breaks down compounds that are not dissolved in the sulfuric acid digestion. The LTIMP THP data have been converted to TP by linear regression (Hatch, 1997).

4.5 Iron

Since the inception of the LTIMP, iron has been measured and reported in five different forms: Dissolved Biologically-available Iron (DBAFe/DFe), Dissolved Iron (DRFe), Soluble Iron (SolFe), Biologically-available Total Iron ((TBAFe/BAFe), and Total Iron (TRFe). The sources of iron in basin streams include weathered mineral particles and groundwater discharge. Iron has been identified as a limiting nutrient in Lake Tahoe, and six streams are 303(d)-listed for Total Iron. Data for Total Iron, however, are only available for the water years 1981-91, with most of the 765 samples concentrated in 1988-90. The Water Quality Objective for Tahoe basin streams is 30 µg/l, but in the entire record, virtually no samples met this objective. The data indicate that Total Iron is strongly influenced by discharge. The natural sources of iron (suspended sediment) overwhelm the anthropogenic sources and the availability of iron is controlled by chelation with organic matter rather than by the supply of Total Iron. A possible mechanism for the role of biological-available iron as a limiting nutrient is its importance in nitrate reduction (Chang et al. 1992). Since the lake is no longer N-limited, iron would seem to be of lesser importance now than in previous years. Total iron is not targeted for control in the Lake Tahoe TMDL plan (Lahontan and NDEP 2010).

4.6 Temperature

Water temperature is routinely measured by USGS and TERC when water samples are collected. In recent years it has been measured at short intervals *in situ* along with pH, conductivity and turbidity as digital recording thermographs have come down in price. Temperature is very important biologically, for its influence on metabolic rate of organisms and on dissolved oxygen concentration. With the coming climatic changes that are anticipated for the Tahoe basin (Coats et al., 2013; Sahoo et al., 2013), good records of stream temperature will become increasingly important.

4.7 Conductivity

Conductivity, like temperature, is routinely measured by USGS at the LTIMP gaging stations. It is closely related to Total Dissolved Solids~~,~~ and, while not by itself an important water quality parameter in basin streams, it may be a valuable ancillary variable for use in modeling concentrations of other constituents.

5 Sampling Strategy and Methodology

LTIMP sampling frequency and protocols have shifted over the years with changes in program management and budgets. There are general guidelines but no formal specification of a protocol. The current sampling program uses three sampling schedules: systematic monthly sampling, intensive storm and intensive snowmelt sampling. Some samples are collected at low as well high flows, but the emphasis is on sampling during high flows (Alvarez, 2006). The average number of samples per year was generally between 60 and 100 from 1973 to 1989, then dropped sharply (after the USGS took over station operations) in 1990, to between 20 and 30 samples per year, where it has remained until today (see Figure 5.1). Since 1990 the west-side stations (WC-8, BC-1, and GC-1) have been sampled the most frequently, averaging 33 to 36 samples per station per year. The west-side stations are operated by Scott Hackley of TERC and include some nighttime sampling. Currently, only 7 stations are being operated: WC-8, BC-1, GC-1, IN-1, TC-1, TH-1, and UT-1. In 2012 (the most recent year in our data set), between 20 and 26 samples were collected at these stations.

Figure 5-1. Mean number of LTIMP water quality samples per station per year. Statistics include only station-years in which at least 10 samples were collected.

Currently, monthly samples are collected during the first week of each month, all year. Spring runoff samples are taken as soon as low elevation snow starts melting – as early as March. During the spring, samples are spaced at least 3 days apart. Spring runoff samples comprise about as many samples as monthly samples, but may be more in bigger runoff years. An effort is made to sample on the day of peak annual runoff. In addition, up to about 5 samples per year are collected during large storms if there is enough warning and it can be done during normal working hours. At the west-side stations more storm samples are collected-- sometimes 2 or 3 during a single event. In general the frequency of sampling is greater during high flows, but is more systematic than random, i.e. not too clustered.

Changes in time of sampling may have introduced bias in calculations of total load. Prior to 1989 USGS staff sampled intensively during snowmelt with samples collected both day and night. Due to (very real) safety concerns nighttime sampling was cut back or discontinued for streams in Nevada as well as Trout Creek and the Upper Truckee River (UTR). The TRG/TERC, however, has continued nighttime sampling on Ward, Blackwood and General Creeks. On the larger watersheds, especially in the latter days of the snowmelt season, the daily snowmelt pulse (which carries most of the daily water volume and constituent loads) may arrive at the gaging station after dark. Eliminating nighttime sampling could thus create an apparent downward trend in total load or concentration (see Section 9).

Robertson and Roerish (1999) simulated different sampling strategies by subsampling water quality data from eight small streams in Wisconsin. They considered sampling at semi-monthly, monthly and 6-week intervals, with and without supplemental sampling during storms, storm peaks, or on rising hydrographs using single-stage samplers (Edwards and Glysson, 1999). Although hydographs in these agricultural environments may be quite different from those in the Tahoe basin, many of the same principles are applicable. They found that the most precise and least biased estimates resulted when multiple years of data from semi-monthly sampling were combined. There can be some loss of precision in combining data from multiple years when regression relationships vary from year to year. The high precision that they reported probably resulted from the large sample sizes that would accrue from such a strategy--large enough to include an adequate representation of high flow events. Strategies that incorporated supplemental sampling during storms were biased because (1) the supplemental samples did not properly represent all parts of the hydrograph and (2) the unbiasedness of the regression estimation method depends on having samples of average concentration for a given discharge. Because concentrations of sediment and associated chemical constituents in the Tahoe basin and elsewhere tend to be greater for a given discharge when the stream is rising, samples should not be preferentially collected during particular parts of the hydrograph. If too many samples are collected during rising hydrographs they will positively bias the estimates, and if they are collected exclusively during falling hydrographs they will negatively bias the estimates. This is particularly a problem in snowmelt environments where hydrographs have a 24-hour period and sampling is only done during the day. It also can be a concern during storm chasing, where crews usually arrive after the peak of the event. It is important to ensure that high flows are sampled, particularly when sample sizes are small, but the best strategies seem to be systematic or random sampling~~,~~ if estimates will be based on regression of concentration versus discharge. Such an approach generally results in the majority of samples on the falling limb of the hydrograph and that is appropriate because the falling limbs are longer in duration than rising limbs. Robertson and Roerish reported that the bias from storm chasing was less than that from peak or single-stage sampling precisely because crews tended to arrive after the peak. In validation of the LTIMP protocols they found that fixed-period monthly sampling supplemented by storm chasing was the most effective strategy. However, traversing stations in a particular order during each storm event is probably not the optimal approach because it will tend to result in certain locations being sampled early in the hydrogaph and others being consistently sampled late.

With the use of automatic pumping samplers, possibly controlled by programmable data loggers, efficient sampling strategies can be implemented that improve precision and accuracy in load estimation. Many more samples can be collected without additional field effort, and well-studied strategies can be implemented via algorithms without regard to time of day or convenience (Thomas, 1985; Thomas and Lewis, 1993; Thomas and Lewis, 1996). These strategies have not been considered for the LTIMP program for various reasons (cost, freezing temperatures, delays in retrieving samples), but pumping samplers have been used with some success for sampling storm runoff in the Tahoe basin. Implementations in which sampling is based on turbidity (Turbidity Threshold Sampling) have greatly improved load estimates in environments where freezing temperatures are not common (Eads and Lewis, 2003). Turbidity is nearly always more closely related to sediment concentration than is discharge. Hysteresis in the relationship is usually minor or nonexistent. The main objective of sampling in a program with continuous turbidity is to cover the full range of concentrations – the time-of sampling bias becomes irrelevant or at least much less important. The same is likely to be true for sediment ~~-~~related chemistry parameters.

Comment [RT1]: Is this true? It would seem to me that, at least for sediment, with higher concentrations on the rising limbs the turbidity would be greater as well.

In the early years of LTIMP~~,~~ chemistry samples were collected as grab samples~~,~~ by filling a hand-held sample bottle from the creek bank. Suspended sediment samples were collected using the equal-width increment method (EWI). In this method, a sample bottle is inserted into the DH-48 sampler and lowered at a uniform rate from the surface to the bottom at equally-spaced intervals across the stream. Since 1988, chemistry samples have been collected by the EWI method. Sampled water is placed in a churn, thoroughly mixed, and subsamples withdrawn for analysis. Separate 1-liter plastic bottles are filled for TKN and TP, and a filtered sample is obtained for dissolved constituents. Figures 5-2(a) through 5-2(f) show the steps in sample collection.

Proper sample storage and handling is very important. Changes (uptake and release) can occur in the biologically ~~-~~active species. Fine sediment is at risk for flocculation, which can change the particle size distribution. The current practice is to store samples in a cooler immediately after collection for same-day delivery to the lab.

Figure 5-2. Scott Hackley of TERC collecting samples at Blackwood Creek. (a) Stretching a tape across the creek for EWI sampling; (b) Collecting a sample with the DH81 sampler; (c) Pouring the sample into the churn splitter. Photos by Jim Markle (http://jimmarkle.smugmug.com/)

Figure 5-2. Scott Hackley of TERC collecting samples at Blackwood Creek. (d) Filtering a sample for analysis of dissolved constituients; (e) Collecting an unfiltered sample from the churn splitter for TP and TKN; (f) Labeling a sample bottle. Photos by Jim Markle (http://jimmarkle.smugmug.com/)

6 Calculating Total Loads

Conceptually, the calculation of tributary mass loads requires evaluating an integral. The load in a given time interval between t_1 and t_2 is given by

$$L = \int_{t_1}^{t_2} K Q_t C_t \, dt$$

where L is the total load in the time interval t_1 to t_2; 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 must be sampled in the field and later determined in a laboratory.

A number of methods and various refinements have been developed for estimating loads. These fall into three categories: Integrating methods, model-based methods, and design-based methods. Integrating methods estimate concentration by averaging successive concentrations or interpolating between them, multiplying by the discharge between sampling times, and summing 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, averaging successive concentrations rather than interpolating between them.. The PWS has been used at Hubbard Brook to calculate total ion loads leaving the watersheds (Likens *et al.*, 1977).

Model-based methods use statistical relationships to predict concentration from other variables, such as discharge or turbidity One type of model-based method is the ratio estimator, which assumes 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 response variable and x_i is the explanatory or predictor variable. This condition is often met with instantaneous load as the response variable and instantaneous discharge as the explanatory variable (Preston *et al.*, 1989). Note that storm-chasing or snowmelt chasing (as used by LTIMP) are not random sampling strategies.

Regression estimators, another model-based method, 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 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). Of course the relationship can also be applied to instantaneous discharge if available, and one would expect better results than at the daily time step. We refer to these methods, whether applied to instantaneous or daily data, as the Rating Curve (RC) Method.

Design-based methods are applicable when samples have been collected using random sampling designs. 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.

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 requires automated sampling equipment and programmable data loggers.

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.

Two methods of load estimation have been used by LTIMP: the worked record and the rating curve method. The method of the worked record, used in the early days of LTIMP, may be thought of as an interpolating 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 intra-daily 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.

Beginning in 1988, a variant of the rating curve method replaced the Worked record method, and has been used since 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 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) and can mislead hydrologists into using models that have no explanatory value for concentration. Figure 6-1 shows the methods used for nutrients and sediment by LTIMP for the primary stations from 1980 through 2012.

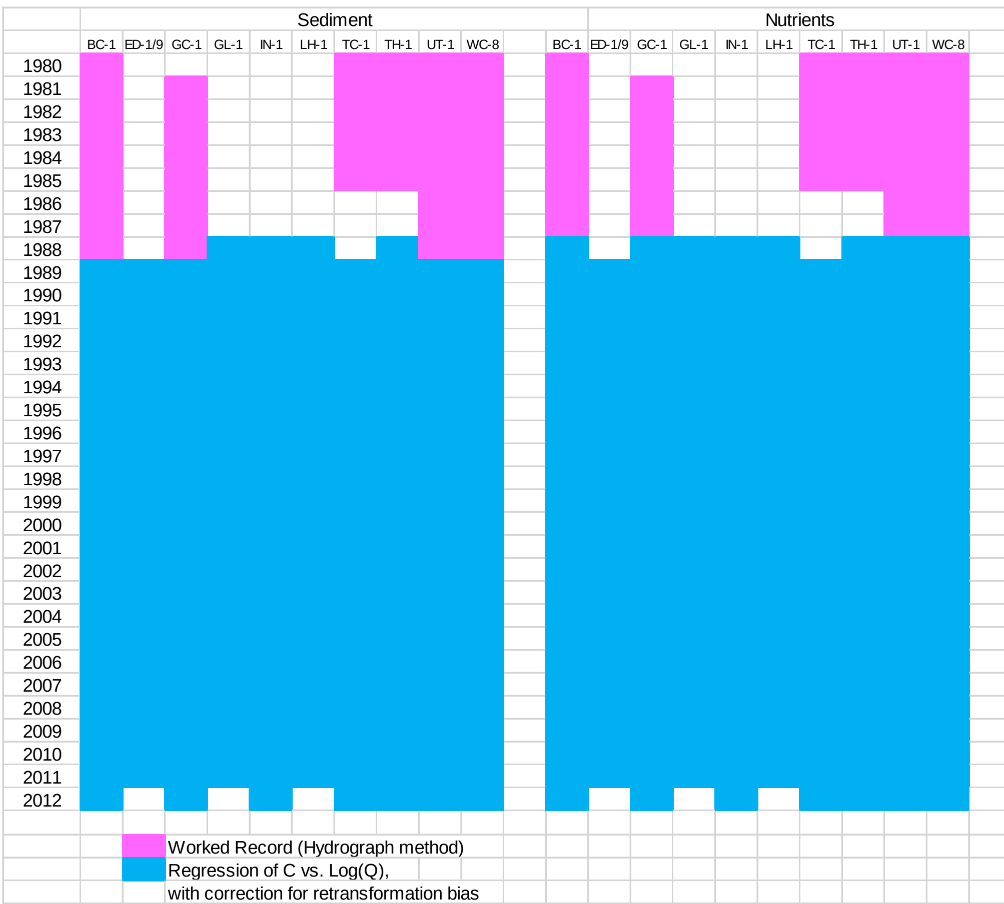


Figure 6-1. Summary of load calculation methods used for LTIMP streams, 1980-2012

7 Statistical Criteria for a Monitoring Program

The two most important properties of a monitoring program are bias and precision. Bias in a monitoring program is the difference between the average of measurements made on the same object (such as total annual load) and its true value. Precision refers to the degree to which repeated measurements under unchanged conditions show the same result. Figure 7-1 illustrates the two concepts.

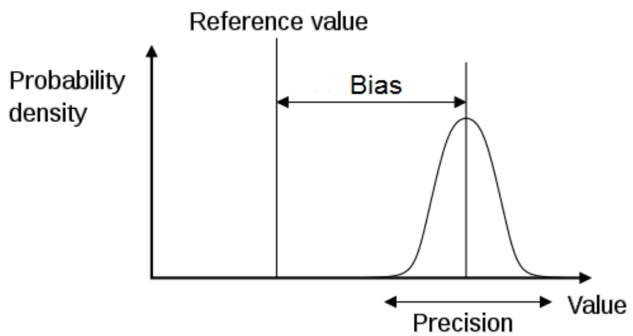


Figure 7-1. Bias and precision in a sampling program

A sampling program can produce results that are unbiased and precise, unbiased but imprecise, biased but precise, or both biased and imprecise. Random error can reduce precision without affecting bias; systematic error increases bias, but not necessarily precision. Accuracy, as we use the term, requires both low bias and high precision, though it has sometimes been used as a synonym for unbiased. Our goal in reviewing the LTIMP is to improve its accuracy by reducing bias and increasing precision, in a cost-effective manner.

In a program as complex as LTIMP, there are many sources of error. It is important that managers, as well as field, laboratory and office personnel understand these sources. Basing load estimates on automated probes (such as turbidity, nitrate, etc.) has the potential to improve both the precision and bias of estimates, but can introduce new sources of error that must be managed. For estimating sediment loads using continuous discharge or turbidity data and possibly with pumped samples, short-term sources of error may include the following:

1. Stage-discharge rating equation
    a. Inadequate data set
        i. Data not representative of full range of discharge

        ii. Data not representative of full time period
        iii. Inadequate number of discharge measurements to estimate rating equation reliably
    b. Bias in electronic stage/or and discharge
        i. Stage measurement device not indexed accurately to staff plate
        ii. Instrument drift (changes in calibration)
        iii. Current meter spins too slow
        iv. Discharge not measured at staff plate location
        v. Errors in stage due to ice formation (see Section 3 above)
    c. Regression model has high variance
        i. Errors in reading staff plate and derivation of mean gage height
        ii. Measurement errors in discharge field procedure or calculations
        iii. Random scour and fill occurred during time period represented by regression data
    d. Regression model is inappropriate for the data
        i. Selected model does not fit entire range of data
        ii. Relationship shifted but only a single model was applied
        iii. Relationship is not linear; has wrong functional form
2. Sediment sampling and estimation
    a. Inadequate data set
        i. Data not representative of full range of SSC
        ii. Data not representative of full time period (esp. when autosampler is out of bottles)
        iii. Inadequate number of samples were collected to reliably estimate relation of SSC to predictor variable(s)
        iv. Samples preferentially collected at certain times of day or on certain parts of the hydrograph, for example on falling limb or at low flows (see Section 10 below).

b. Bias in SSC and predictor variables
    i. Unrepresentative pump intake location if SSC not spatially uniform
    ii. Inadequate pumping sampler line velocity for coarse particles
    iii. Algal growth from delays in processing unrefrigerated samples
    iv. Turbidity or stage sensor drift (calibration changes)

c. Regression model has high variance (can be statistically estimated)
    i. Pump sample malfunctions
        1. Inadequate purge cycle (residual sediment left in pump tubing leading to cross-contamination between samples)
        2. Overfilled bottles
    ii. Turbidity measurement errors
        1. Minor biofouling, not corrected
        2. Interpolation and reconstruction errors
        3. Changes in T-probe location if turbidity not spatially uniform
    iii. SSC (lab) measurement errors
        1. Spillage of sediment or sample
        2. Weighing errors
        3. Bookkeeping errors (e.g. mislabeled samples)
        4. Calculation errors

d. Regression model is inappropriate for the data
    i. Selected model does not fit entire range of the data
    ii. Relationship shifted but only a single model was applied
    iii. Relationship is not linear; has wrong functional form
    iv. Model is based on external data due to lack of samples

Sources of error for chemical methods are discussed above in Section 4.

8 Statistical Tests of Sampling and Load Calculation Methods

A major objective of this project was to determine the most accurate methods for calculating historic and future loads. Based on Monte Carlo sampling of worked records from Blackwood Creek (1985-1986), Coats et al., (2002) recommended the PWS method for estimating nitrate and SRP loads. For SS and TP, the preferred method differed between 1985 and 1986. The RC method was superior for SS in 1986 (the wetter year) but was very biased for SS loads in 1985. For TP, the RC methods were recommended when regressions of concentration on discharge are statistically significant, otherwise the PWS method should be used.

We decided to run additional Monte Carlo simulations using a wider variety of populations and estimation methods. These simulations involve resampling from (1) synthetic populations of target variables or (2) worked records developed from stations and years when sample sizes were large. The synthetic data sets were constructed using regression relationships with transformed and untransformed continuous variables (discharge, lagged discharge, turbidity, conductivity, water temperature) from previous studies. Random error was added to the synthetic concentrations. For maximum relevance to historic load estimation, the data sets were resampled in a way that retains the characteristics of historic LTIMP sampling protocols, and loads are estimated using multiple methodologies on each synthesized sample. The worked records were resampled by subsampling the LTIMP water quality data.

For each simulation, we recorded the precision, bias, root mean squared error (RMSE), and the 50th, 80th, 90th, and 95th percentiles of the absolute value of the error. The 50th percentile is synonymous with median absolute percent error (MAPE). The best methods were judged using two measures of accuracy: RMSE and MAPE. The latter is less sensitive to occasional extreme errors. For each population (synthetic data set or worked record), methods were ranked for each sample size and then ordered by the sum of ranks over all sample sizes. The top 6 methods by sum of ranks were plotted for each population. To resolve differences in results among different sampled populations, RMSE or MAPE for each method and sample size was averaged across populations before ranking.

8.1 Synthetic data sets

In order to compare different sampling designs and load calculation models, we created realistic synthetic data sets of the major water quality constituents of concern: SSC, FS, TP, SRP, TKN, and NO_3-N, and resampled these sets up to 90 times in each test of a method. To create the synthetic data sets, we obtained data sets from several researchers who have conducted studies in sub-basins around the lake: Trout Creek (Gary Conley, 2nd Nature LLC), Homewood Creek (Mark Grismer, 2013), Angora Creek (Alan Heyvaert), Ward and Blackwood Creeks (Andrew Stubblefield), Rosewood and Third Creeks (Rick Susfalk). An essential feature of all these data sets is that they included near-continuous turbidity data that could be correlated with sampled constituent data to produce near-continuous concentration data. In addition to the turbidity data, several stations included near-continuous conductivity or water temperature data. All these variables, together with discharge, gave us the ability to develop models for realistically constructing synthetic data sets that we could use for simulating sampling protocols and load estimation methods.

The data sets vary in quality and some of them needed extensive cleaning up to remove sunspikes, irregular intervals, gaps, timestamp errors, daylight savings changes, and other problems in the data. The discharge and turbidity data were cleaned using the TTS Adjuster graphical user interface developed at Redwood Sciences Laboratory. Data quality and incompleteness of the turbidity data limited the number and length of useable data sets. Only Homewood Creek 2010 had an entire water year of acceptable data, so partial water years were used for most of the synthetic data sets (Table 8.1-1). All regression models for SRP (Angora, Third, Trout, and Ward Creeks) and NO_3-N (Angora, Third, Trout, and Homewood Creeks) exhibited relatively high variance. We wanted to use the most realistic populations available, so we relied on the worked records rather than synthetic data for simulating SRP and NO3.

Table 8.1-1. Synthetic data sets used for simulations

Watershed	***LTIMP ID***	***Year***	***Period***	***# Days***	***SSC***	***FS***	***FSP***	***TP***	***TKN***
Angora	AC-2	2008	Mar18-Jun09	84				x	x
Angora	AC-2	2010	Apr17-Sep30	167			x		
Homewood	none	2010	Oct01-Sep30	365	x	x		x	x
Rosewood	none	2010	Oct01-May31	243		x			
Third	TH-1	2005	Feb24-Jun22	119	x				
Trout Pioneer	TC-2	2010	May13-Jul08	57	x				
Trout Pioneer	TC-2	2011	Apr05-Jul30	117	x			x	
Trout Martin	TC-4	2010	Apr16-Jul08	84		x			
Trout R1	none	2011	Apr05-Aug10	128	x				
Ward	WC-8	1999	May13-Jun24	43	x			x	x

FS – fine sediment mass concentration (mg/L <16 or <20 microns)
FSP – fine sediment particle concentration (particles/ml)

The turbidity (T) data were most useful for modeling SSC, FS, TKN, and TP. Other predictor variables entertained for modeling the various water quality parameters were instantaneous discharge (Q), discharge lagged by h (h=3,6,12,24) hours (Q_h), mean daily discharge (MDQ), mean daily discharge lagged by d (d=1,2) days (MDQ_d), number of days since start of water year (D), conductivity (C), and water temperature (WT). The lagged discharge variables were included for their potential to model hysteresis, wherein concentrations are greater for a given discharge during the rising portion of a daily or seasonal hydrograph. In some cases (e.g. Trout Creek models for SSC) data were combined from multiple locations in a watershed and/or multiple water years at the same location in order to obtain a robust model. In such cases water year was retained in the model as a categorical variable if it was a statistically significant predictor. Location was not a significant predictor in any of these models.

Regression predictors discussed above, and their logarithms, were screened in the R computing environment (R Development Core Team (2010), using all-subsets-regression in the *leaps* package (Lumley and Miller, 2009). The best models with 1, 2, and 3 predictors were identified for each response and its logarithm. The overall best generating model (Table 8.1-2) was selected through a combination of significance testing on predictors and evaluation of model diagnostics, especially linearity of predictors, normality of residuals, and homogeneity of variance.

Table 8.1-2. Models used to generate synthetic data sets

Watershed	*Year*	*Model*	*AR1*	*R^2*	*RSE*	*Load*
Trout TC-2	2010	$SSC = 0.482T^{1.459} + 1.283\log(Q)$	0.50	0.961	6.76	310ton
Trout TC-2	2011	$SSC = 0.482T^{1.459} + 1.283\log(Q)$	0.50	0.961	6.76	310ton
Trout R1	2011	$SSC = 0.482T^{1.459} + 1.283\log(Q)$	0.50	0.961	6.76	310ton
Ward	1999	$SSC = 1.377T^{1.153}$	0.50	0.925	9.38	320ton
Homewood	2010	$SSC^{0.5} = 0.531 + 1.16T^{0.5}$	0.66	0.974	0.40	549kg
Third	2008	$SSC^{0.5} = 0.991 + 1.50T^{0.5}$	0.66	0.778	1.50	113ton
Homewood	2010	$FS = 0.239T^{1.086}$	0.50	0.997	0.32	95kg
Rosewood	2010	$FS = 119 + 0.979T$	0.98	0.864	50.0	28ton
Trout TC-4	2010	$FS^{0.5} = 0.722 + 0.107T + 0.243\log(Q)$	0.95	0.823	0.47	51ton
Angora	2010	$\log(FSP) = 4.832 + 0.648\log(MDQ) - 0.173MDQ2 + 1.658C$	0.95	0.821	0.47	3.1×10^{16} particles
Angora	2008	$TP = 1.712 + 1.943T^{0.5} + 1.021\log(Q) - 0.911\log(Q24)$	0.99	0.847	0.92	28kg
Trout TC-2	2011	$TP = 15.70 + 4.499T + 0.469Q - 0.490MDQ2$	0.95	0.814	6.06	1032kg
Ward	1999	$TP = 0.5838 + 3.995T$	0.95	0.951	6.37	628kg
Homewood	2010	$\log(TP) = 4.140 + 0.343T^{0.5} - 0.00674D$	0.95	0.894	0.28	1.4kg
Angora	2008	$\log(TKN) = 4.822 + 0.449\log(Q) + 0.383\log(T) - 0.367\log(MDQ2)$	0.999	0.691	0.45	445kg
Ward	1999	$\log(TKN) = 8.360 + 0.898\log(T) - 1.050\log(Q)$	0.98	0.767	0.38	2.20ton
Homewood	2010	$\log(TKN) = 5.926 + 0.457\log(T) - 0.00524D$	0.998	0.732	0.55	12kg

AR1 – autoregressive order 1 parameter, used to generate autocorrelated random errors
R^2 – multiple coefficient of determination
RSE – residual standard error

Once the models were identified, they were applied to the continuous data sets at 30-minute intervals to generate synthetic traces. If one of these populations were to be sampled in a simulation, and a model of the same form as the generating model was fitted to the data, the model would fit the data perfectly with no error. Therefore it was necessary to add random error to the predicted concentrations. Because of the short 30-min measurement interval in these data sets, deviations from the modeled concentration values would almost certainly be highly autocorrelated. Sets of high-frequency measurements of concentration needed for modeling autocorrelated errors are exceedingly rare. The only such data known to the authors are 10-minute records of turbidity and SSC from several storm events at the Caspar Creek Experimental Watershed (Lewis, 1996). At 10-minute intervals, residuals from regression of SSC (or transformed SSC) on turbidity exhibited autocorrelation explainable by autoregressive (AR) models up to order 3. At 30-minute intervals, AR models of order 1 were adequate to explain the serial correlation. The exact AR coefficient varied by storm event and depended also on the SSC transformation. One of the Caspar Creek storm events had considerably more data than the others and the AR order 1 coefficients from this event (for modeling SSC, log(SSC) and $SSC^{0.5}$) were used as starting points for representing the Tahoe basin errors. For each synthetic data set, autoregressive random error was generated using the appropriate coefficient and rescaled so its variance matched the variance of the generating model. Of course Caspar Creek is quite different from the Tahoe basins and, in any case, there is no reason to think these coefficients would be appropriate for parameters other than SSC. An interactive tool with a slider bar was developed to visualize the synthetic trace for any chosen AR coefficient. In many cases the traces appeared too noisy with the Caspar SSC coefficients, so the coefficients were made closer to 1 in order to smooth out the traces while keeping the overall error variance fixed. Figures 8.1-1 and 8.1-2 show the effects of increasing the AR coefficient from 0.5 to 0.98 for a portion of the record at station TC-2. In this case, synthetic TP was generated from a model involving both turbidity and discharge (Table 8.1-2).

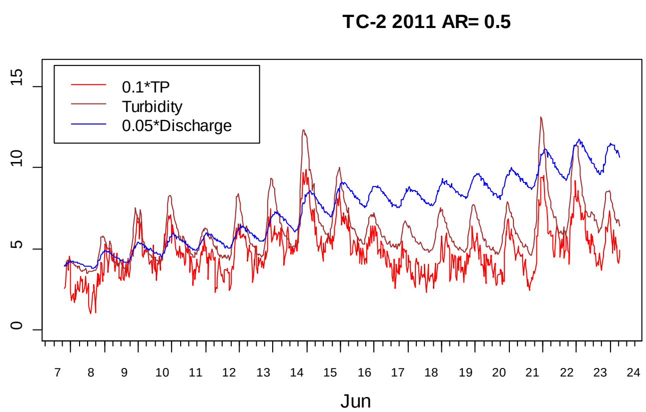


Figure 8.1-1 . Synthetic TP trace with Caspar Creek autoregressive coefficient, 0.5.

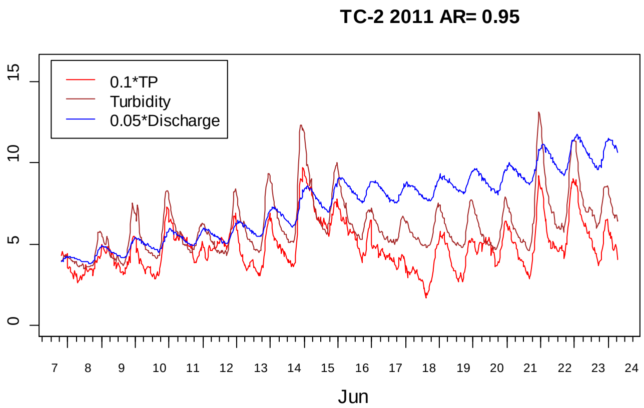


Figure 8.1-2 . Synthetic TP trace smoothed by increasing autoregressive coefficient to 0.95.

8.1.1 Sampling from the synthetic records

To create a grid of sampling times, each interval between actual LTIMP samples was subdivided into equal subintervals, preserving the relative sampling frequencies at different times of the year, e.g. least frequent during off-season, more frequent during snowmelt, and most frequent during storm-chasing. For sample sizes of 10, 20, 40, 60, and 80, an appropriately-spaced grid of sampling times was created and all possible systematic samples of that size were sampled from the grid (Figure 8.1-3).

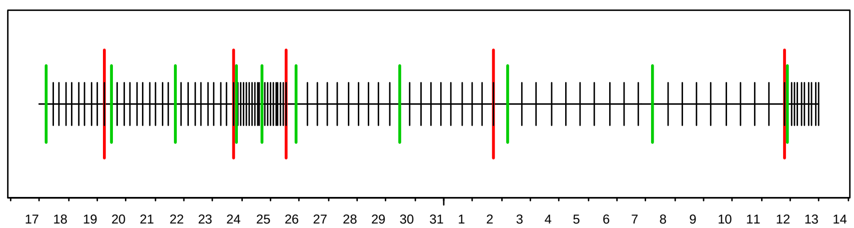


Figure 8.1-3. Illustration of algorithm for sampling synthetic data sets. Axis ticks denote days. Red segments show actual LTIMP sample times and constitute a sampling "template". Black segments delineate equal subintervals between each pair of red segments. Green segments show a systematic sample of the black segments. Simulations generate all possible systematic samples from a given set of subintervals.

Finally, to make sampling more realistic, in half the simulations selected times were mapped from a 24-hour day to a 9-6pm workday (Figure 8.1-4). Sampling on weekends and holidays was permitted.

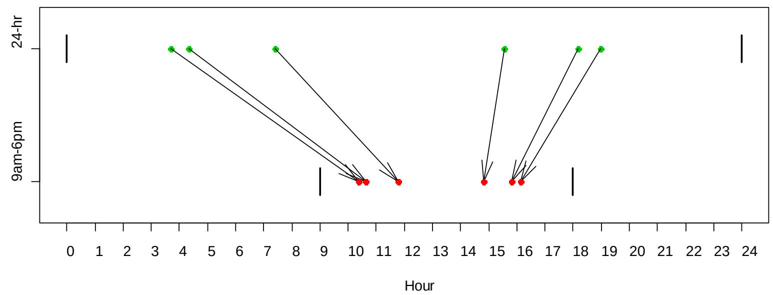


Figure 8.1-4. Illustration of mapping the 24-hour day onto a 9am-6pm work day.

The number of points (times) per subdivided interval is *npi = nsim*samsize/n0*, where *nsim*=number of simulated samples desired, *samsize*=simulated sample size, and *n0*=the actual LTIMP sample size. An additional *npi* points are also inserted before the first LTIMP sample time and after the last LTIMP sample time, half of them before and half after. In order to get a constant sample size, points are in turn discarded in approximately equal numbers from the start and end of the sampling grid to reduce the grid size to exactly *nsim*samsize*. The value of *nsim* was kept at 30 for all simulations.

To boost the number of simulated samples we used 3 sets of actual LTIMP times for each simulation, from stations that are on substantially different sampling schedules (Figure 8.1-5). In most cases we had to insert or omit one or two sampling times to get the same *n0* for all three stations. The result was 90 simulated samples for each combination of station and year.

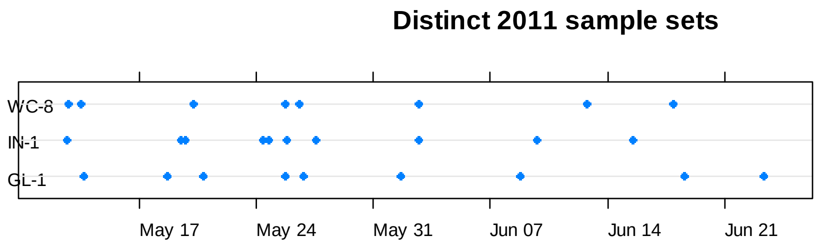


Figure 8.1-5. Three sampling templates used to sample synthetic data for WC-8 in 2011

8.2 The worked records

The method of the worked record was described earlier in section 6. The worked records consist of mean daily flows and mean daily concentrations derived from sample data and hand-drawn concentration curves. Because of the relatively large sample sizes, some of these worked records provide a fairly detailed representation of transport that is useful for Monte Carlo simulations. The worked records include 7 gaging stations: BC-1, GC-1, SN-1, TC-1, TH-1, UT-1, and WC-8. These records were compiled for years 1976-78 and 1980-1987, a total of 11 years, but the records are not complete for some combinations of constituents and years. We limited the simulations to combinations that were complete from April 15 to July 15 (Table 8.2-1), however if that period was included in a longer complete period, the entire period was used. In the majority of cases, the entire year (365 or 366 days) was used (Table 8.2-2).

Table 8.2-1. Number of worked records with complete data from Apr 15 to Jul 15

	BC-1	*GC-1*	*SN-1*	*TC-1*	*TH-1*	*UT-1*	*WC-8*	**Total**
NO3	11	7	5	6	6	7	11	53
SRP	10	7	5	6	6	7	10	51
SSC	9	7	5	6	2	7	9	45
THP	7	6	4	4	4	7	7	39

Table 8.2-2. Number of worked records with complete data from Oct 1 to Sep 30

	BC-1	*GC-1*	*SN-1*	*TC-1*	*TH-1*	*UT-1*	*WC-8*	**Total**
NO3	6	6	4	4	2	6	6	34
SRP	6	6	4	4	4	6	6	36
SSC	6	6	4	4	4	6	6	36
THP	6	6	4	4	4	6	6	36

Worked records are available for NO3, SRP, SSC, and THP (total hydrolyzable phosphorus). Figure 8.2-6 is an example worked record for NO3. Simulations were done on all 4 constituents, but the SSC and TP populations derived from turbidity are more realistic than the worked records and we favor the methods indicated by the continuous-data simulations.

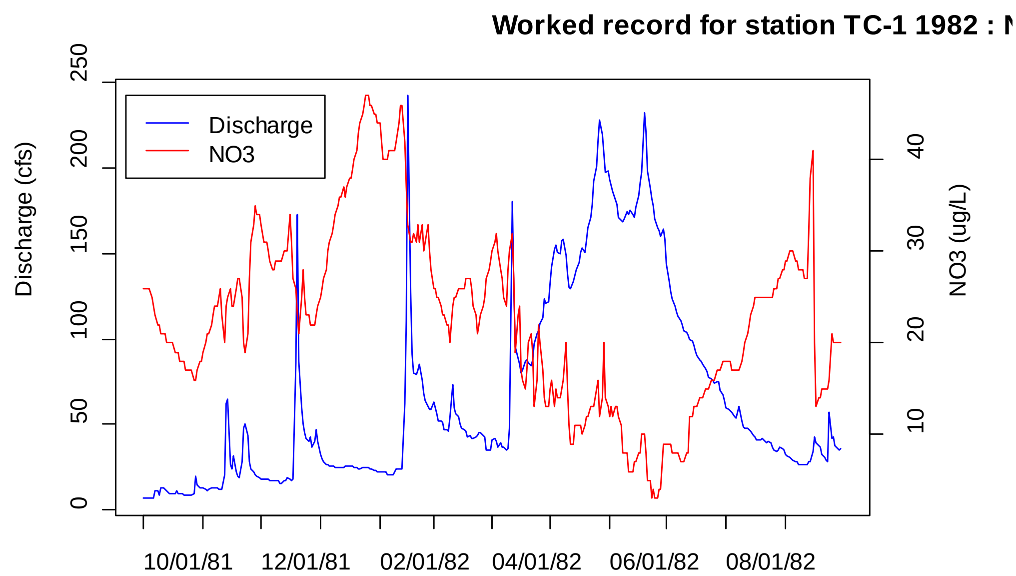


Figure 8.2-6. Example of a worked record for nitrate

8.2.1 Sampling from the worked records

The sampling algorithm used on the continuous data sets was not practical for the worked records because of their low (daily) time resolution. Days actually sampled by LTIMP were subsampled to obtain simulation sample sizes of 10, 20, 40, 60, and 80. Subsampling was not possible when the LTIMP sample size was smaller than the desired simulation sample size (Table 8.2-3). Otherwise, subsampling was done 10 times for each simulation sample size, and results from all combinations of station and year were pooled. The simulated sample sizes were often a large fraction of the sample sizes used to develop the worked record. In these cases, primarily for simulated sample sizes of 60 or more, it is likely that (1) the worked record simulations underestimate operational errors, and (2) simulations would tend to favor the period-weighted sampling estimator, because it is similar to the interpolating methods used to derive the worked records.

Table 8.2-3. LTIMP sample sizes for each worked record in Table 8.2-1

		1976	1977	1978	1980	1981	1982	1983	1984	1985	1986	1987
NO3	BC-1	78	30	80	123	91	95	91	96	79	86	68
	GC-1					50	84	86	78	73	83	58
	SN-1					96	74	80	66	51		
	TC-1				75	88	118	85	70	80		
	TH-1				102	108	92	91	101	86		
	UT-1					90	113	92	71	81	81	57
	WC-8	81	32	80	123	91	97	89	95	83	87	68
SRP	BC-1		26	80	122	95	95	91	96	79	86	68
	GC-1					50	84	86	78	73	83	58
	SN-1					96	74	80	66	51		
	TC-1				75	88	118	85	70	80		
	TH-1				102	108	92	91	101	86		
	UT-1					90	113	92	71	81	81	57
	WC-8		28	80	123	91	97	89	95	83	87	68
SSC	BC-1	78			123	95	95	91	96	79	86	68
	GC-1					50	84	86	78	73	83	58
	SN-1					96	74	80	66	51		
	TC-1				74	88	118	85	70	80		
	TH-1								101	86		
	UT-1					90	113	92	71	81	81	57
	WC-8	81			123	91	97	89	95	83	87	68
THP	BC-1					82	95	91	96	79	86	68
	GC-1						84	86	78	73	83	58
	SN-1						74	80	66	51		
	TC-1						118	85	70	80		
	TH-1						92	91	101	86		
	UT-1					74	113	92	71	81	81	57
	WC-8					82	97	89	95	83	87	68

Many of the worked records for SSC contain concentrations of zero (Table 8.2-4). These samples were omitted from models using logarithms of SSC concentration, since it is impossible to take the logarithm of zero. In sampling from regression-generated synthetic populations, we excluded all values of SSC < 0.5 to avoid extremely small values of log(SSC) that would exert too much leverage on the regression lines. This was not an issue with the worked records, which do not contain *non-zero* values smaller than 1 mg/L.

Table 8.2-4. Percentages of zeroes in each worked record in Table 8.2-1

	BC-1	*GC-1*	*SN-1*	*TC-1*	*TH-1*	*UT-1*	**WC-8**
NO3	0	0	0	0	0	0	0
SRP	0.8	0	0	0	0	0	0
SSC	7.2	27.2	1.4	0	0	1.3	17.8
THP	0	0	0	0	0	0	0

Comment [RT2]: I thought it might be well to emphasize this since I first read this sentence as a contradiction of the first sentence in this paragraph.

8.3 Load estimation methods

For the Monte Carlo simulations the load was estimated using a variety of regression, both parametric and non-parametric, and interpolating methods (Table 8.3-1). The variables used in the regression models are a subset of those introduced in section 8.1.3 including turbidity (T), instantaneous discharge (Q), discharge lagged by 24 hours (Q_{24}), mean daily discharge (MDQ), mean daily discharge lagged by 1 day (MDQ_1), and number of days since start of water year (D). Most of the regression models used a log-transformed response and log-transformations for all the predictor variables except D. Two of the regression models used square root transformations for the response and predictor T or Q. A few of the models utilized turbidity or discharge at 30-minute intervals to predict a high resolution time series of concentration; however, since neither turbidity nor continuous discharge are available historically, most of the models utilized daily mean discharge in the predictions. Simulations from the worked records (i.e. those for SRP and NO3) used only models that did not require turbidity and that could be applied at a daily time step.

In addition to the specific parametric regression models, procedures were developed to select the best regression model among those that used a log-transformed response. Four model selection methods (rcb1, rcb2, rcb3, and rcb4) were compared, using (see Glossary) Akaike's Information Criterion (AIC) and Gilroy's (Gilroy et al., 1990) estimate of root mean square error (GRMSE) for the predicted load. Each of the model selection criteria, AIC_c and GRMSE were used to select among (1) those models for log(C) that do not depend on turbidity as a predictor (rcb1 and rcb2), or (2) the full set of log(C) models (rcb3 and rcb4); see Table 8.3-1.

Several non-parametric methods were implemented that model and predict concentration from discharge. These methods create best-fit curves of arbitrary shape using statistical algorithms, and include (see Glossary):

- LOcally-wEighted Scatterplot Smoothing (Cleveland and Devlin, 1988) also known as a LOESS smoothing
- Alternating Conditional Expectations (ACE)
- Additivity and VAriance Stabilization for regression (AVAS)

Table 8.3-1. Load estimation methods used in simulations

Short name	Type	Time step	Description
rcload.turb	simple regression	30-min	log(C) ~ log(T)
rcload.turb2	simple regression	30-min	$C^{0.5} \sim T^{0.5}$
rcload	simple regression	30-min	log(C) ~ log(Q)
rcload.mdq	simple regression	daily	log(C) ~ log(Q)
rcload.mdq2	multiple regression	daily	log(C) ~ log(Q) + log(MDQ/MDQ_1)
rcload.mdq3	multiple regression	daily	log(C) ~ log(Q) + log(MDQ_1)
rcload.mdq4	multiple regression	daily	log(C) ~ log(Q) + log(MDQ/MDQ_1)+D
rcload.mdq5	multiple regression	daily	log(C) ~ log(Q) + D
rcload.mdq6	simple regression	daily	$C^{0.5} \sim Q^{0.5}$
loess.g	nonparametric	daily	Loess with gaussian fitting
loess.s	nonparametric	daily	Loess with symmetric fitting
ace	nonparametric	daily	ACE transformations (see text)
avas	nonparametric	daily	AVAS transformations (see text)
areg.boot	nonparametric	daily	AVAS *areg.boot* implementation
pdmean	averaging	daily	Period-weighted sampling estimator
pdlinear	interpolating	daily	Daily interpolator
pdinstant	interpolating	30-min	“Continuous” interpolator
pdlocal2	interpolating	daily	Two-point rating curves + global curve
pdlocal2a	interpolating	daily	Two-point rating curves + interpolation
pdlocal4	interpolating	daily	Four-point rating curves + global curve
rcb1	best regression	daily	Selection by AIC without turbidity
rcb2	best regression	daily	Selection by GRMSE without turbidity
rcb3	best regression	daily	Selection by AIC, all log(C) models
rcb4	best regression	daily	Selection by GRMSE, all log(C) models

The averaging and interpolating methods that we implemented include

- The period-weighted sampling estimator (pdmean) was described in Section 6. 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 daily interpolator (pdlinear) is similar to pdmean, but concentration is interpolated at a daily time step. The load is computed as the sum of the products of daily mean discharge and daily concentration.
- The continuous interpolator (pdinstant) is like pdlinear but operates at the time-step of the original measurements. For our synthetic data sets, the time-step is 30 minutes.
- Two-point rating curves utilize 2-point discharge rating curves in place of linear interpolation to interpolate daily concentrations. The rating curve for the segment between each pair of samples was computed from the concentrations and discharges of

just those two samples. Because these rating curves are often ill-conditioned, they are only utilized for segments in which the exponent lies within a specified positive range. For other segments, one of two methods was employed: (1) a global rating curve is utilized (pdlocal2), or (2) linear interpolation as in the pdlinear method (pdlocal2a).

- Four-point rating curves (pdlocal4) work very similarly to the two-point methods but the sample size is boosted from 2 to 4 by including the adjacent samples before and after the segment being estimated. For ill-conditioned segments a global rating curve was utilized; the linear interpolation method was not implemented for pdlocal4.

While the list of load estimation methods we have tested may seem extensive, it is by no means complete. There are many more models that have potential; limited resources prevented us from testing them all. Classes of models that deserve further exploration are:

- Power models. These models appear to be equivalent algebraically to logarthmic models but they are solved in original units without taking logarithms, using non-linear least squares. Computations can be carried out without eliminating zeroes from the data set and the solutions give less weight to the typically abundant small values that become highly dispersed when transformed by logarithms.
- Mixed turbidity/discharge models. Regression models employing both turbidity and discharge were useful in creating some of the synthetic data sets, and presumably could be useful in an operational mode.
- More functional forms for turbidity models. We only considered two forms of turbidity models: linear and square-root. Other forms, such as logarithmic, power, and loess models have proven very useful in estimating sediment loads at gaging stations where turbidity is measured continuously.

8.4 Evaluation of optimal handling of low SSC values

When the logarithm of concentration is computed for log-regression models, small values less than 1 become widely dispersed (Figure 8.4-1) even though the practical differences among them are insignificant. This is particularly a problem with suspended sediment, whose concentration is often very low or below detection limits. Regression variance is typically very high in the lower ranges of discharge or concentration, and the overall relationships may be non-linear. Because they are so numerous and widely spread in logarithmic space, measurements at low concentrations and discharges can exert a great deal of influence (i.e. leverage) on the regression line, causing underestimation or overestimation of concentrations at the upper end of the range. We evaluated two options for handling low SSC values below specified limits of 0.10, 0.25, 0.50, and 1.00 mg/L: (1) discarding and (2) setting to half the limit. For stations and years in which worked records existed and LTIMP data contained sample concentrations less than 1.0 mg/L, we compared regression estimates using method rcb2 based on the LTIMP samples to the loads that had been computed from the worked records. Estimates closer to that from the worked record are considered more accurate.

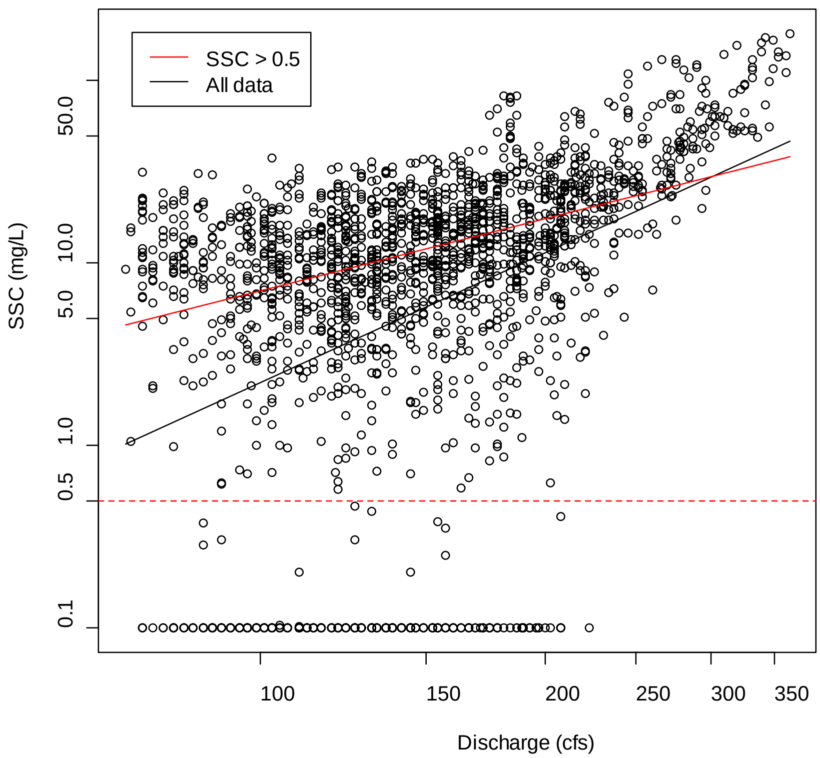


Figure 8.4-1. Synthetic data for WC-8 WY1999. Values less than 0.1 mg/L were set to 0.1 to allow computation of logarithms from samples. Relationships between SSC and discharge are shown for (1) data pairs with SSC > 0.5, and (2) all data.

Table 8.4-1. Sediment loads computed from worked records and estimated from LTIMP samples using regression method rcb2 with various rules for handling small values of SSC. Highlighted cells show the regression estimate closest to the worked record.

		Worked	Sample	Discard values less than				Set to half if less than			
Site	Year	Record	Size	1.00	0.50	0.25	0.10	1.00	0.50	0.25	0.10
GC-1	1982	57.98	114	78.42	76.62	83.76	82.49	93.36	92.71	85.46	81.58
GC-1	1983	12.05	107	13.90	13.07	13.39	13.61	14.69	14.57	13.78	13.61
GC-1	1984	7.37	100	7.08	6.93	6.85	6.87	7.17	7.17	7.05	6.90
GC-1	1985	2.34	85	2.64	2.52	2.51	2.51	2.58	2.58	2.51	2.51
GC-1	1986	27.81	128	27.81	24.82	25.28	25.55	28.73	28.16	26.68	26.07
GC-1	1987	0.88	70	1.34	1.05	1.05	1.04	1.13	1.11	1.07	1.06
WC-8	1982	269.97	129	307.81	298.13	321.14	321.14	347.55	335.39	321.14	321.14
WC-8	1983	55.84	110	59.07	59.20	59.60	60.32	61.42	60.91	60.91	60.32
WC-8	1985	4.15	102	4.13	3.98	4.01	4.04	4.24	4.22	4.09	4.04
WC-8	1986	57.85	136	53.64	53.69	48.65	49.29	54.87	52.67	49.55	49.29
WC-8	1987	2.23	92	5.82	5.16	4.53	4.52	4.68	4.68	4.56	4.52
BC-1	1982	285.23	115	494.11	541.15	651.12	651.12	774.50	766.17	705.01	680.79
BC-1	1983	81.81	111	94.54	99.11	101.54	101.54	104.43	103.79	101.54	101.54
BC-1	1984	39.06	124	48.30	48.45	48.70	48.70	49.13	49.13	48.70	48.70
BC-1	1985	11.56	94	12.62	12.50	12.64	12.88	13.40	13.29	13.03	12.88

Discarding low values nearly always produced more accurate estimates than setting these values to half the detection limit. Setting the MDL to either 0.5 or 1.0 and discarding lower values was the best method in 11 of 15 tests, with MDL=1.0 winning in 6 of those cases. All the exceptions were years with very small sediment loads. Since the data set did not contained only values between 0.5 and 1.0, discarding values less than 1.0 was equivalent to discarding values of 0.5 or less, and that is the method we settled upon for the simulations.

Comment [RT3]: Sentence not clear. You must mean "Since the data sets did not contain any values between…;" right? But if this is true why aren't the loads computed for the two discard levels the same in table 8.4-1? Also, the "that" reference for "that is the method" is a bit murky.

8.5 Simulation results for daytime sampling

Figure 8.5-1 shows the 6 top-ranking methods by RMSE in the simulation of suspended sediment load estimation using the synthetic data for station WC-8 in WY 1999. At all sample sizes the best (most accurate) method is rcload.turb2, the square root regression model for turbidity. The result is expected because that was~~is~~ the model~~ that was~~ used to generate the synthetic data set. The next best model selection methods with turbidity were ~~(~~rcb3 and rcb4~~)~~ ~~were the next best methods~~; recall that these methods are restricted to searching models for log(C), which was not optimal for this population. Turbidity is not available for load estimation with historic data. The best method without turbidity was LOESS with gaussian fitting (loess.g), although it was tied with some other methods at sample sizes of 10, 60 and 80.

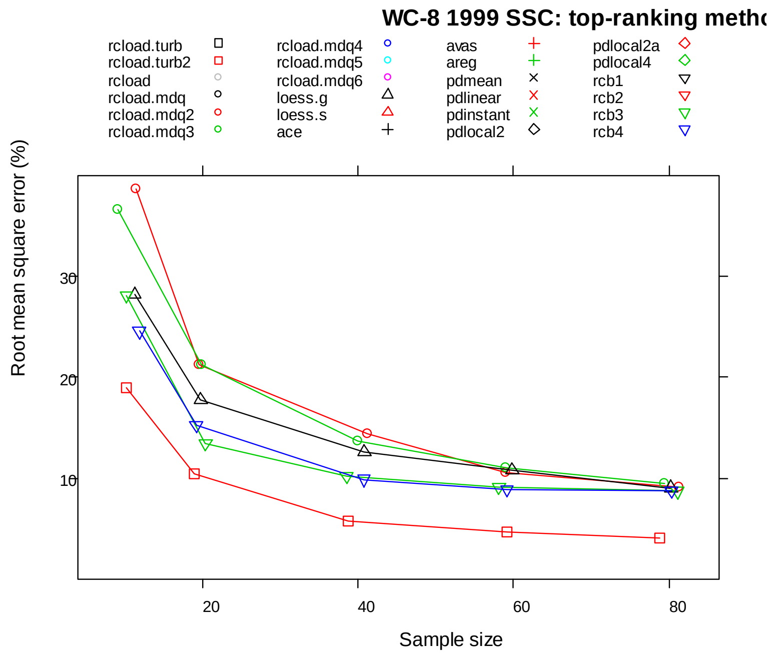


Figure 8.5-1. The 6 top-ranking methods by RMSE in the simulation of suspended sediment sampling from the synthetic data for station WC-8 in WY 1999.

Note that the simple rating curve (rcload.mdq) was not among the top-ranking methods in this simulation. The most accurate method depends on both the population that is being sampled, sample size, and the selection criterion (Tables 8.5-1 and 8.5-2).

Table 8.5-1. Best estimation method for suspended sediment according to either RMSE or MAPE from simulated sampling of synthetic data sets. Methods requiring turbidity are included. Last two rows show method with lowest mean rank for given sample size.

Station	year	Criterion	n=10	n=20	n=40	n=60	n=80
WC-8	1999	RMSE	rcload.turb2	rcload.turb2	rcload.turb2	rcload.turb2	rcload.turb2
		MAPE	rcload.turb2	rcload.turb2	rcload.turb2	rcload.turb2	rcload.turb2
TC-R1	2011	RMSE	pdlocal4	pdlocal4	rcload.mdq6	rcload.mdq6	rcload.mdq6
		MAPE	pdlocal4	pdlocal2a	pdlocal2	rcload.mdq6	rcload.mdq6
HWD	2010	RMSE	rcload.turb2	rcload.turb2	rcload.turb2	rcload.turb2	rcload.turb2
		MAPE	rcload.turb2	rcb3	rcload.turb2	rcload.turb2	rcload.turb2
TC-2	2011	RMSE	rcload.mdq	rcload.mdq	rcload.mdq	rcload.turb2	rcload.mdq3
		MAPE	pdlocal2a	rcload.turb2	rcb3	rcload.mdq3	rcload.mdq3
TC-2	2010	RMSE	rcload.mdq	rcload.turb	rcload.mdq3	rcload.mdq3	rcload.mdq3
		MAPE	ace	rcload.mdq3	rcload.mdq3	rcb3	rcload.mdq3
TH-1	2005	RMSE	rcload.turb2	rcload.turb2	rcload.turb2	rcload.turb2	rcload.turb2
		MAPE	rcload.turb2	rcload.turb2	rcload.turb2	rcb3/4	rcb3/4
Combined		RMSE	rcb4	rcload.turb2	rcload.turb2	rcload.turb2	rcload.turb2
		MAPE	rcload.turb2	rcload.turb2	rcb3	rcload.turb2	rcload.turb2

Table 8.5-2. Best estimation method for suspended sediment according to either RMSE or MAPE from simulated sampling of synthetic data sets. Methods requiring turbidity are omitted. Last two rows show method with lowest mean rank for given sample size.

Station	year	Criterion	n=10	n=20	n=40	n=60	n=80
WC-8	1999	RMSE	loess.g	loess.g	loess.g	rcload.mdq2	pdmean
		MAPE	rcb1	loess.g	loess.g	areg	pdlinear
TC-R1	2011	RMSE	pdlocal4	pdlocal4	rcload.mdq6	rcload.mdq6	rcload.mdq6
		MAPE	pdlocal4	pdlocal2a	pdlocal2	rcload.mdq6	rcload.mdq6
HWD	2010	RMSE	rcb2	rcb2	rcb2	rcb2	rcb2
		MAPE	rcb2	rcb2	rcload.mdq3	rcload.mdq2	rcload.mdq3
TC-2	2011	RMSE	rcload.mdq	rcload.mdq	rcload.mdq	rcload.mdq3	rcload.mdq3
		MAPE	pdlocal2a	pdlocal2a	rcload.mdq2	rcload.mdq3	rcload.mdq3
TC-2	2010	RMSE	rcload.mdq	rcload.mdq	rcload.mdq	rcload.mdq3	rcload.mdq3
		MAPE	ace	rcload.mdq3	rcload.mdq3	rcload.mdq3	rcload.mdq3
TH-1	2005	RMSE	pdlocal2a	pdmean	pdlocal2a	pdlocal4	loess.s
		MAPE	pdlocal4	pdlocal2a	pdlocal4	pdlocal4	pdlocal4
Combined		RMSE	pdlocal2a	rcload.mdq2	rcb2	rcload.mdq2	rcload.mdq2
		MAPE	rcload.mdq	pdlocal2a	rcload.mdq2	rcload.mdq2	rcload.mdq3

The turbidity models performed better and there was more consistency of results when turbidity was available, but the latter may be partly due to the fact that the number of methods simulated for turbidity was fewer. Figure 8.5-2 shows the 6 top-ranking methods for suspended sediment based on the mean RMSE across populations and sample sizes. We see that, as for WC-8, rcload.turb2 and other turbidity methods excel, but when turbidity is excluded, the best method is rcb2, the best model selection method based on GRMSE. For suspended sediment, selection by MAPE produced similar rankings as by RMSE.

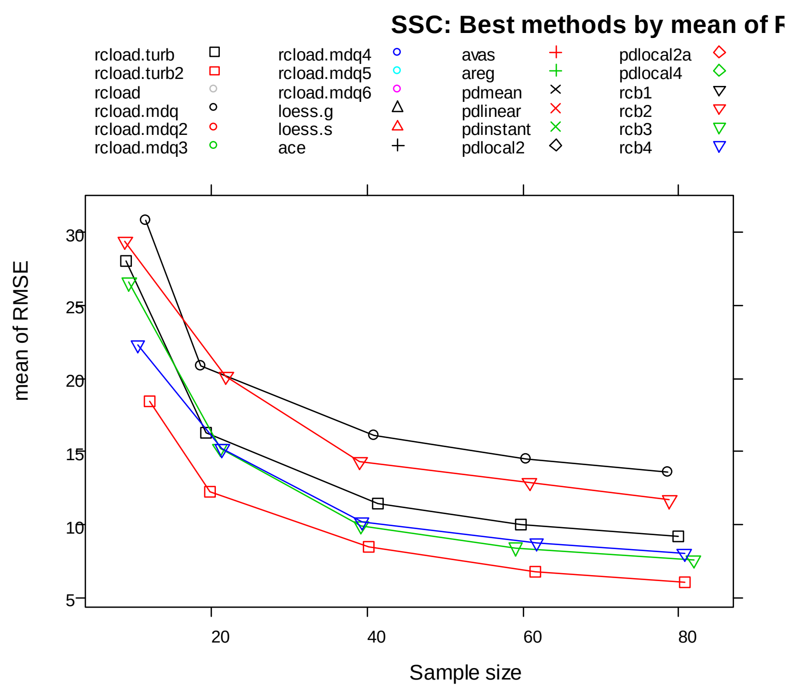


Figure 8.5-2. The 6 top-ranking methods by RMSE in the simulation of suspended sediment sampling from the synthetic data. Rankings are determined by the mean of RMSE (%), first across populations and then across sample sizes.

The dependency of methods on population, sample size, and selection criterion was general across all constituents. Appendix A-2 shows detailed results for all simulations. There is no method that is consistently more accurate than all the rest. The best approach to estimating sediment loads is to have a statistician look carefully at alternative models and make case-by-case decisions that consider goodness-of-fit and adherence to model assumptions. However, for estimating historic loads that will be impractical because we have 5 water quality constituents at

Comment [RT4]: “criteria”?

20 gaging stations over a 43 year period. In the future a case-by-case approach might be possible, but if estimation must be done by a technician without a strong background in statistics, we have selected a few methods based on the overall rankings by mean of RMSE and mean of MAPE (Table 8.5-3). We did not want to recommend a hodgepodge of methods with no logical pattern. Therefore. in a few cases based on other considerations, we selected a method that was nearly the best.

Table 8.5-3. Selected best estimation methods for all constituents with and without turbidity data.

Constituent	***With turbidity***	**Without turbidity**
SSC	rcload.turb2	rcb2
FS	rcload.turb2	rcb2
FSP	rcload.turb2	rcb2
TP	rcb3	rcb2
TKN	pdmean	pdmean
NO3	pdmean	pdmean
SRP	pdmean	pdmean

For example, for estimating TP without turbidity, the best method in the overall rankings was usually rcload.mdq3 or rcload.mdq5, depending on sample size. However the best model varied by synthetic population and we have no assurance that our synthetic populations represent the full range of situations that might be encountered in the Tahoe Basin. Best regression model selection using GRMSE (rcb2) was the method we selected for SS and FS when turbidity was unavailable and the method should do reasonably well with unknown populations because of its flexibility. The method was also one of the three top-ranking methods for TP, so we are recommending rcb2 for TP as well as SSC and FS~~,~~ when turbidity is unavailable. When turbidity is available the best method for TP seems to be rcb3; it slightly outperformed rcload.turb at larger sample sizes and, for unexplained reasons, it did considerably better than rcb4 at all sample sizes.

For fine sediment by mass, rcload.turb2 was the overall best performing method. For fine sediment by particle count, the best method was a standard rating curve (rcload.mdq). However, owing to the paucity of particle count data in the basin~~,~~ we had only one synthetic population of counts (TC-2 2010)~~,~~ and, while turbidity data were available, they were of very poor quality. Theoretically, the count data should be well-correlated to turbidity, and we have assumed for Table 8.5-3 that the most accurate methods for fine sediment mass will also be optimal for fine sediment particle counts.

For TKN, the turbidity methods were best for the Ward Creek and Homewood Creek synthetic data, but were inaccurate for Angora Creek. Averaging/interpolating methods pdmean and pdlocal2a performed well (and similarly to each other) in the overall rankings. In general, the period-weighted sampling method (pdmean, aka PWS) performed very competitively with all of its sister methods (pdlinear, pdinstant, pdlocal2, pdlocal2a, pdlocal4) in all the simulations. While it was not always the best, it was probably the most consistently accurate. Therefore, pdmean was selected as the generally preferred method for estimating TKN load, with or without turbidity. In some specific cases, however, one of the turbidity methods is likely to be more accurate.

For both NO3 and SRP, the worked record simulations indicated that pdmean, pdlinear, and pdlocal2a were the overall most accurate estimation methods. We recommend pdmean for its overall consistency.

The worked record simulations for SS and THP also favored the averaging and interpolating methods. Because the worked records were created using an interpolating process they tend to favor these methods. However, this process may not be the best representation of reality so we believe the results from sampling the synthetic data sets are a better indication of likely performance.

8.6 Recalculated annual loads and total ann. Q

The selected methods (Table 8.5-3) were then used to recalculate annual loads for all stations, years, and constituents. Loads were only calculated when a minimum of 10 samples was available for analysis. For each station, year, and water quality parameter, we saved the following information,which is provided with this report in electronic form as “LTIMP loads.xls”.

- Annual yield computed from all samples collected
- Annual flow volume
- Annual peak daily flow
- Sample size
- Discharge weighted mean concentration
- In the case of best regression, the form of the best model
- Gilroy’s estimate of root mean squared error (GRMSE)
- Annual yield computed from daytime samples only (9am-6pm)
- Daytime sample size

9 Time-of-Sampling bias—methods and results

Because snowmelt occurs preferentially during the day, there is a diurnal periodicity to spring runoff in the Tahoe basin. Daytime sampling therefore systematically oversamples certain parts of the hydrograph and neglects others, introducing a bias in estimating the loads of constituents whose concentration depends on flow. The phasing of diurnal hydrographs depends on factors such as the distribution of elevation and aspect in a basin, size of the basin, and time of year.

In 1988, the USGS cut back on nighttime sampling for safety reasons, but TERC continued with some nighttime sampling in Ward, Blackwood and General Creeks. Figure 9-1 shows the distributions of sampling times for west-side streams and all others, for the earlier and later time periods. The histograms show that there has been some reduction in nighttime sampling in Ward, Blackwood and General Creeks, but a severe reduction for the other streams. The problem of introduced bias due to the change in sampling regimes must be addressed.

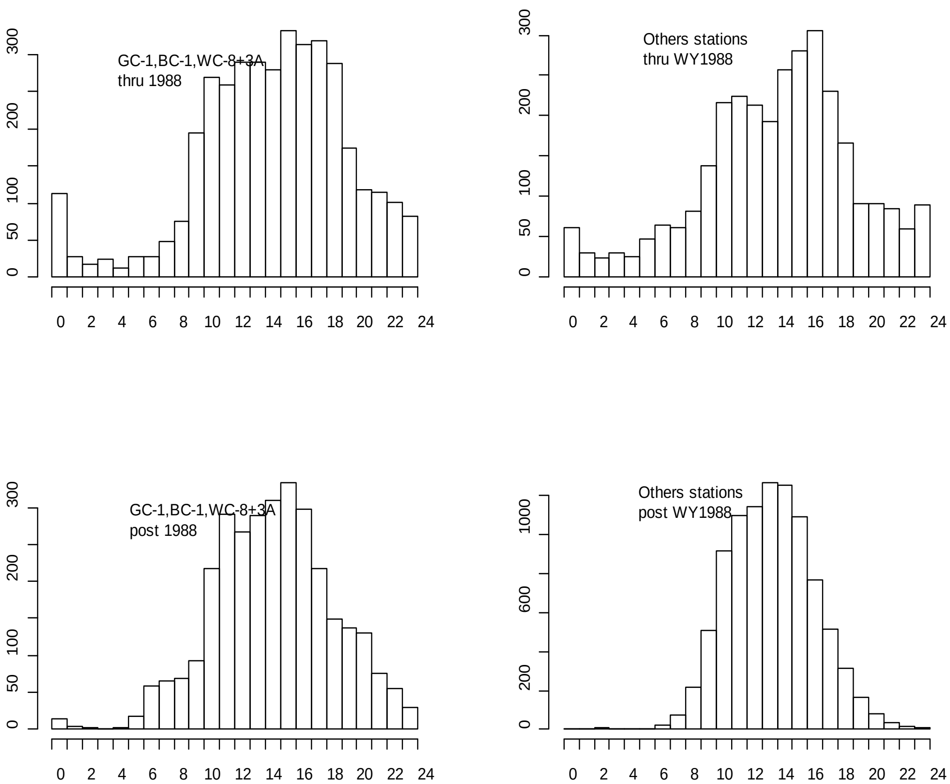
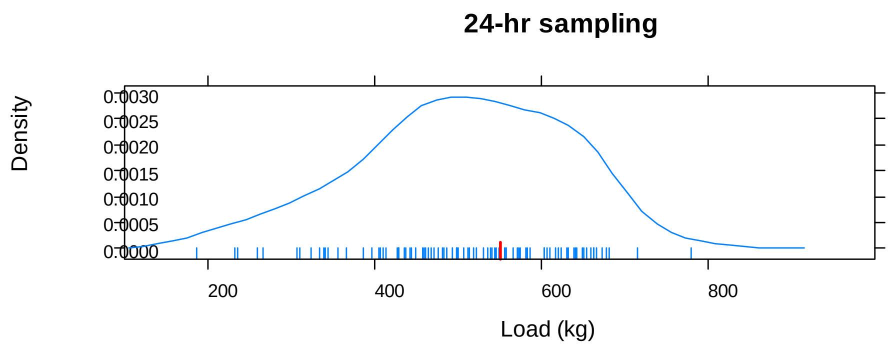


Figure 9-1. Histograms of the distribution of sampling times for the three west-side streams and other streams, for two time periods.

Figure 9-2 shows the distribution of simulated SS load estimates based on a standard log-log sediment:discharge rating curve applied to mean daily discharge, and using the actual LTIMP sample size of n=28. The true synthetic load is shown in both plots as the red bar at 549 kg. The bias in the simulation for 24-hr sampling is -8.1% of the true load (RMSE=23.7%) and that for workday sampling is 11.8% (RMSE=26.5%). The hydrograph generally peaks between 6 and 7pm at Homewood Creek, so daytime sampling essentially restricts samples to the rising limb of the hydrograph when discharge is on a daily snowmelt cycle. Clockwise hysteresis is common in these sediment:discharge relationships (Stubblefield, 2007), explaining why daytime sampling would produce higher estimates than 24-hour sampling.



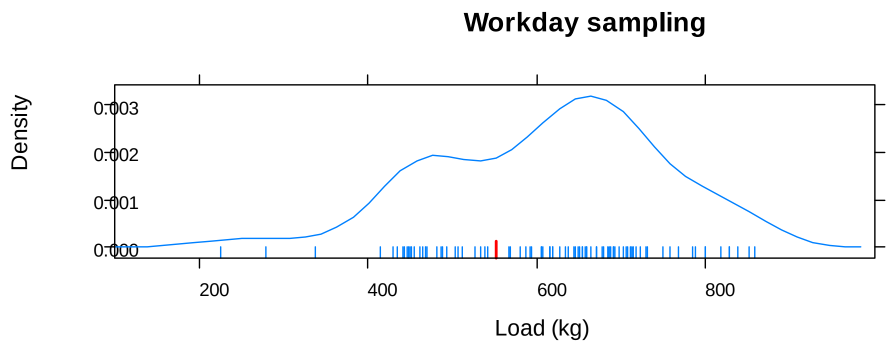


Figure 9-2. Results of 90 simulations at Homewood Creek WY2010 for sediment rating curve load estimates (n=28). Top: 24-hour sampling. Bottom: sampling is limited to a 9am to 6pm workday.

We investigated the time-of-sampling bias using 3 different approaches:

1. We simulated sampling from our synthetic data sets first on a 24-hour basis, and then on a workday (9am-6pm) basis. Both sets of loads were compared with the known loads of these synthetic data sets to determine the bias of both sampling schedules.
2. We calculated historic LTIMP loads first using all samples, and then using only samples collected between 9am and 6pm. We then plotted the relative difference between the loads as a function of number of samples omitted.

3. We tried to induce trends at the west-side stations (GC-1, BC-1, and WC-8) by eliminating night samples from 1992-2012. These are the stations that have the most night-time samples. This approach was intended to reveal the potential importance of the time-of-sampling bias in relation to trend detection; the methods are for this analysis are presented later, in Section 10.

9.1 Simulations

We ran simulations of our selected estimation methods for 24-hour sampling and workday (9am-6pm) sampling for sample sizes of 10 to 80 with SSC, TP, and TKN. The graphs (Figures 9.1-1 - 9.1-3) show the bias as a percentage of true load. The difference between the two curves is the time-of-sampling bias. Not surprisingly, sample size doesn't seem to have much influence on the results. The time of sampling bias is in the same direction for all constituents at a given station, but the direction varies by station. The only station with two synthetic populations (SSC for TC210 and TC211), has a negative time-of-sampling bias in both years but the magnitude differs. At Trout and Angora Creeks the time-of-sampling bias is negative for all constituents (but very small for SSC at TCR11). Hydrographs at all the Trout Creek stations peak after midnight; workday sampling is mostly on the recession limb with a little on the very early rising limb. Hydrograph timing at Angora Creek was very different than at Trout, with peaks in the late afternoon in March 2010 to early evening in May, so it is unexpected that daytime sampling would introduce a negative bias. Ward Creek, Homewood, and Third Creek peaked at 6-8pm and workday sampling is mostly of the rising limb. As expected, the bias was positive at Ward, Homewood, and Third Creeks (Figure 9.1-1). For TKN the time-of-sampling bias is very small for 2 of the 3 populations (Figure 9.1-3). For Homewood 2010, there is a large positive time-of-sampling bias for SSC, TP, and TKN. The time of sampling bias sometimes reduces the overall bias when its sign is opposite that of the bias of the 24-hour method; the prime example is SSC at Homewood 2010.

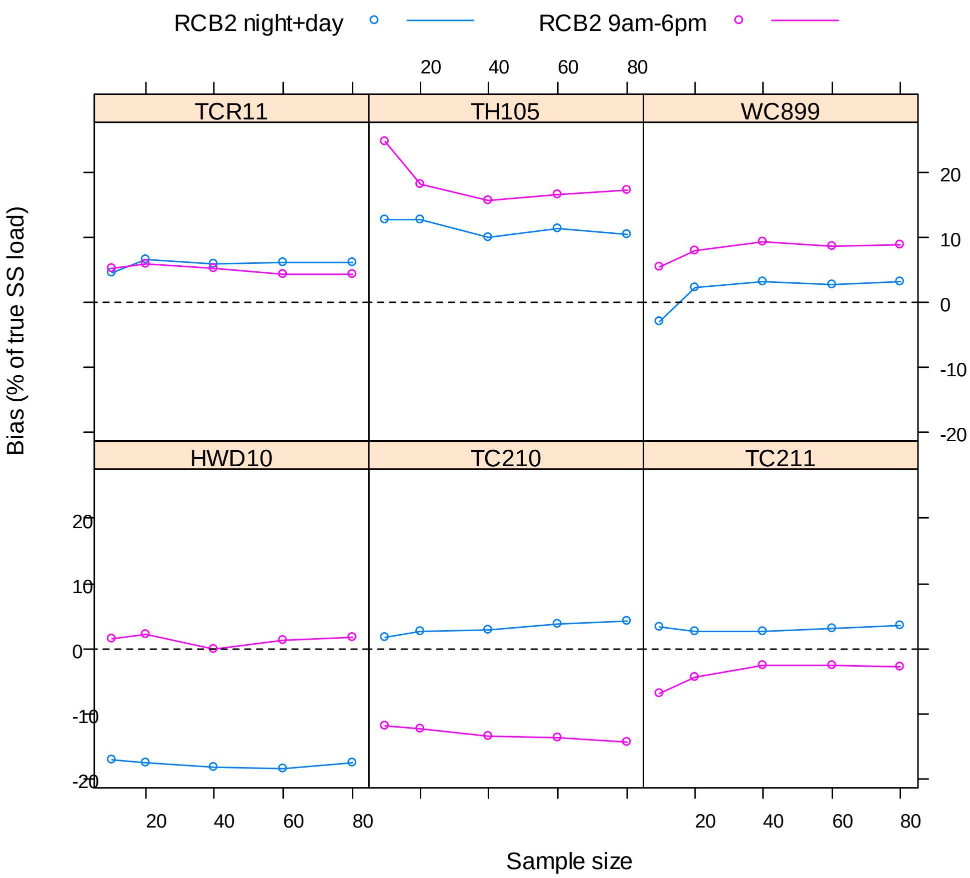
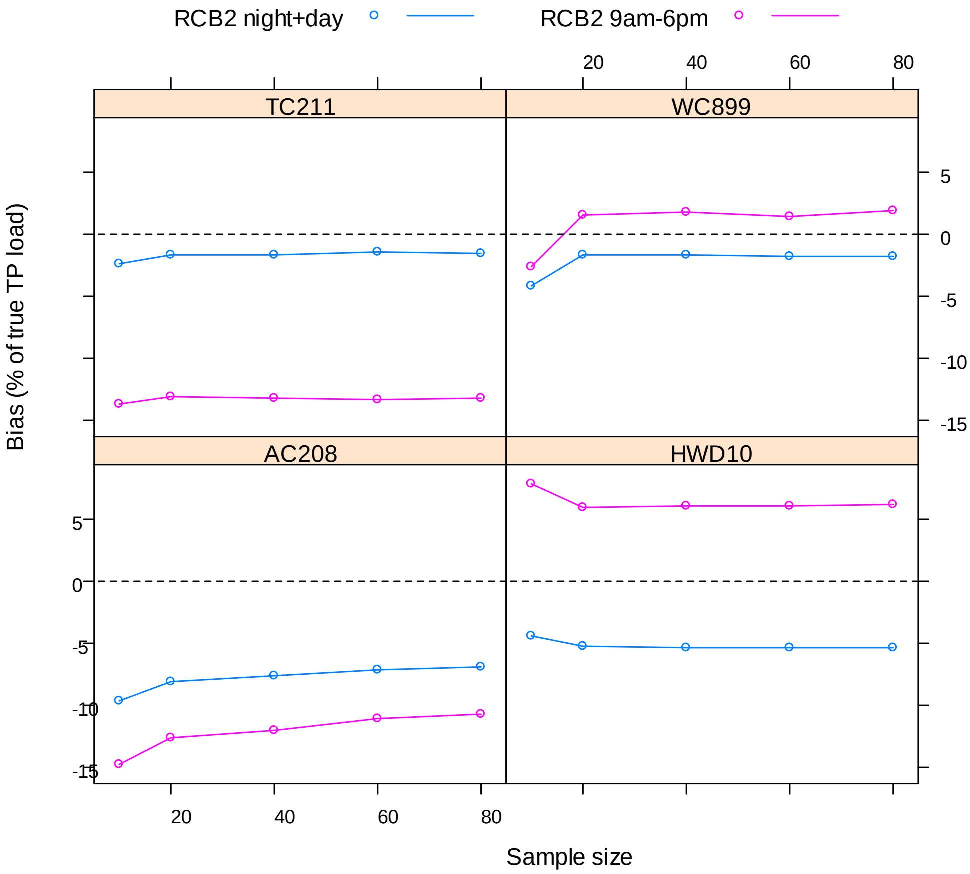


Figure 9.1-1. Bias of the estimation method rcb2 for SS loads, with and without night-time samples



1803
1804 Figure 9.1-2. Bias of the estimation method rcb2 for TP loads, with and without night-time samples
1805

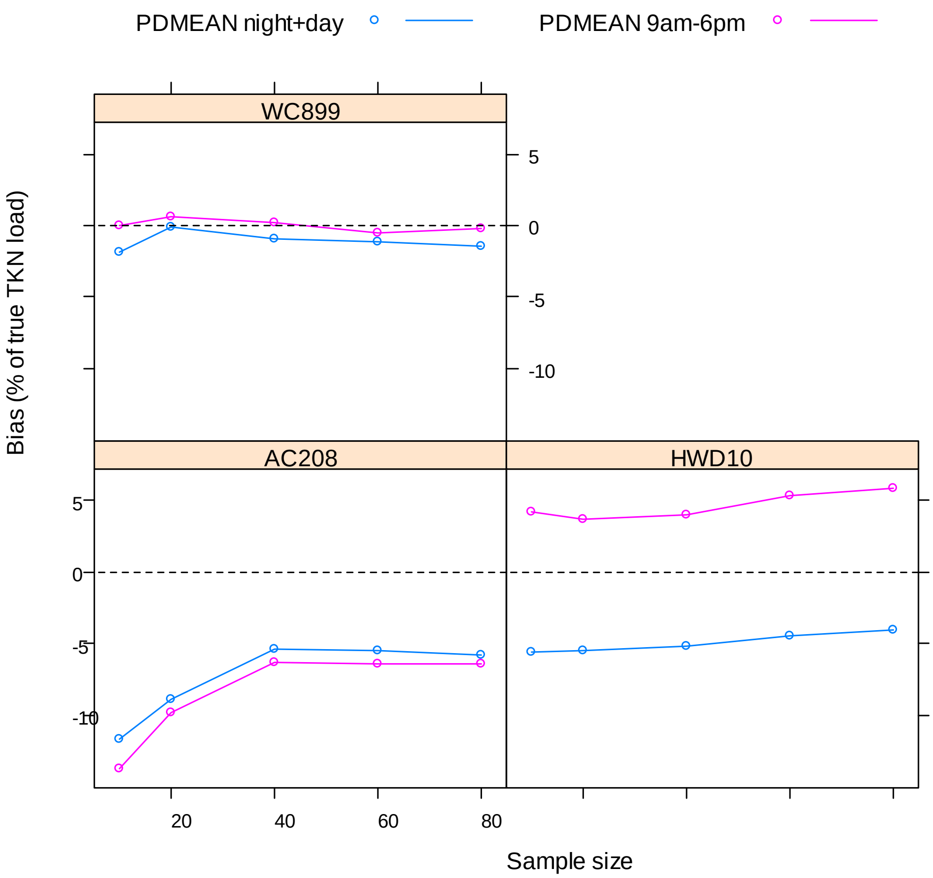


Figure 9.1-3. Bias of the estimation method pdmean for TKN loads, with and without night-time samples

9.2 Historic load estimates

Loads were recomputed for NO3, SRP, TKN, SSC, and TP using only samples collected between 9am and 6pm. Graphs in Appendix A-3 show the change in load as a function of the proportion of samples omitted. We screened results for stations and years where day-only sampling reduced the number of samples by 25% or more, and then averaged over years. The inter-annual variance is large and the number of cases is different for different watersheds. SSC and TP are the most affected. The bias is mostly negative, and generally not large for most stations (Figure 9.2-1), but is probably underestimated because the baseline for comparison in each case is the actual LTIMP sample, which still under-represents nighttime conditions.

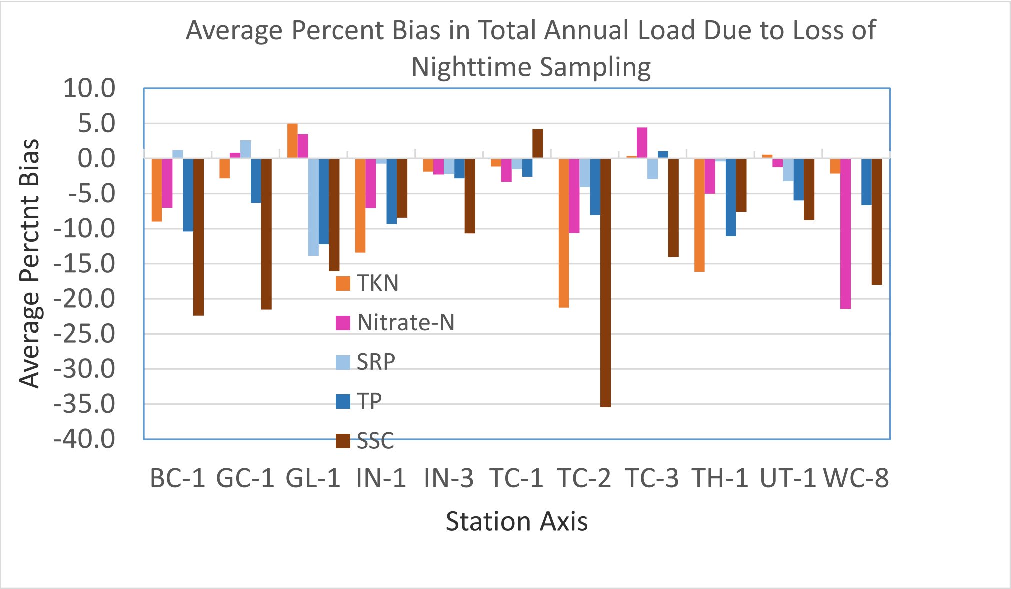


Figure 9.2-1. Average percent change in load from omitting nighttime samples, for those stations and years where this procedure reduced number of samples by 25% or more.

9.3 Induced trends from eliminating night-time samples

To test whether the time-of-sampling bias has the potential to introduce significant false trends in the data, we removed nighttime samples from the data at the west lake stations BC-1, GC-1, and WC-8 *only* for the years 1992-2012. These are stations that historically have been sampled regularly at night. We tested trends only for SSC, TKN, and SRP as trends had already been detected at NO3 and TP after eliminating *all* nighttime samples. We left significance levels for testing at 0.0025 (Bonferroni with n=20 because this was part of a 20-station experiment).

For SS, a decrease in slope is visible at all three stations and a significant trend was induced at WC-8 (Figure 9.3-1 and Table 9.3-1). For TKN, no trends or changes in trend were apparent (Table 9.3-1). For SRP, there were no consequential differences in results with and without night sampling; the SRP trend at GC-1 was highly significant in both cases (Table 9.3-1).

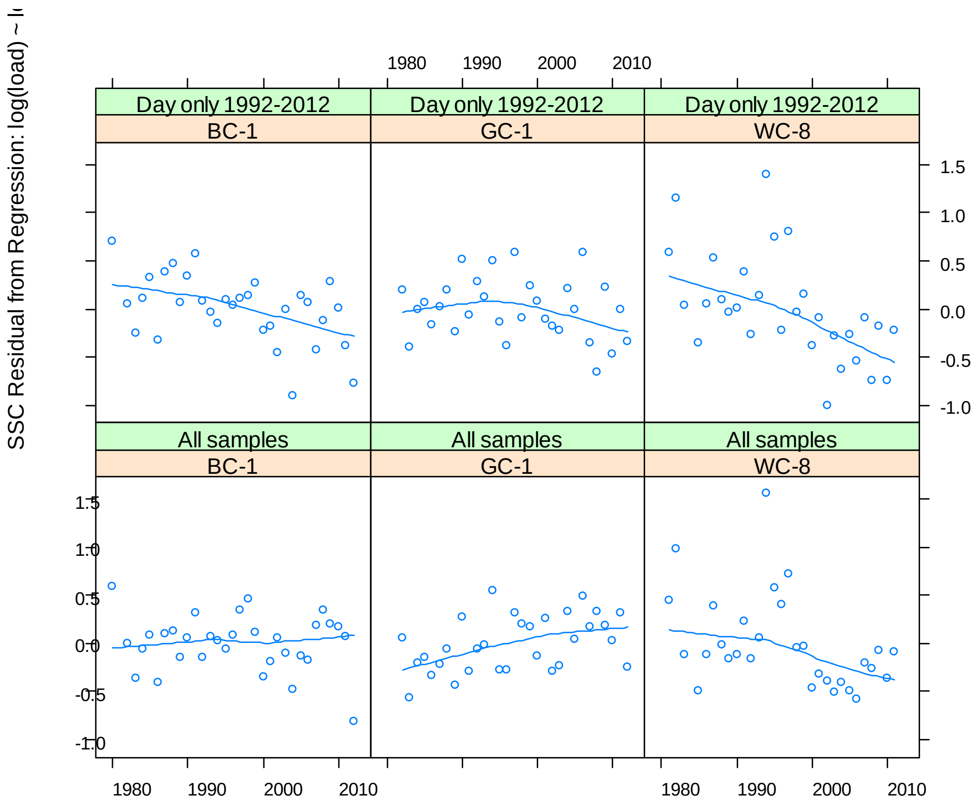


Figure 9.3-1. Trends in SS load with and without including night samples for the period 1992-2012.

Table 9.3-1. Tests of induced trend at west-side stations BC-1, GC-1, and WC-8 for selected constituents. Loads for the years 1992-2012 were computed two ways: (1) with daytime-only samples and (2) with all samples. Table shows the p-values, considered significant (*) when p < 0.0025.

Constituent	Night samples included	BC-1	GC-1	WC-8
SS	No	0.0135	0.3799	0.0007*
SS	Yes	1.0000	0.0199	0.0143
TKN	No	0.6410	0.9804	0.6013
TKN	Yes	0.4461	0.9024	0.5653
SRP	No	0.0176	0.0001*	0.0911
SRP	Yes	0.0233	0.0001*	0.0401

In future sampling of basin streams it will be important to eliminate the time-of-sampling bias Short of sending crews out at night to sample flooding streams in the dark, there are two ways to get unbiased data for the flow-driven constituents (TP, SSC and FS): 1) with pumped samplers and collection of samples around-the-clock; and 2) simultaneous turbidity measurement and sample collection, with extension of the concentration records by regression to cover unsampled time periods.

10 Bias in standard sediment rating curves and optimal methods

The simulations permit a comparison of the bias of the standard rating curve method (rcload.mdq) and the methods we have selected as "optimal". For TKN, TP, and SS, the data come from simulations of the synthetic populations that we developed. These are simulations of daytime sampling (9am-6pm). The "Best Model" by GRMSE (rcb2) method excludes models that require turbidity. For SRP, NO3, and THP, the data come from the worked record simulations, again without the benefit of turbidity. Both rcload.mdq and rcb2 employ Duan's smearing correction for retransformation bias. In all but one simulation, the mean rating curve estimates are greater than or similar to those of our selected method (Appendix A-4.1). The only exception encountered was the simulation for TKN at AC-2 in WY 2008 (Figure 10-1). For most populations and sites we looked at, the selected optimal method is less biased than the standard rating curve, sometimes dramatically so. But there are exceptions wherein our method is more biased: 2 out of 4 for TP, 2 out of 6 for SS. This happens when the selected method in negatively biased. Three of the four cases were at one station (TC-2).

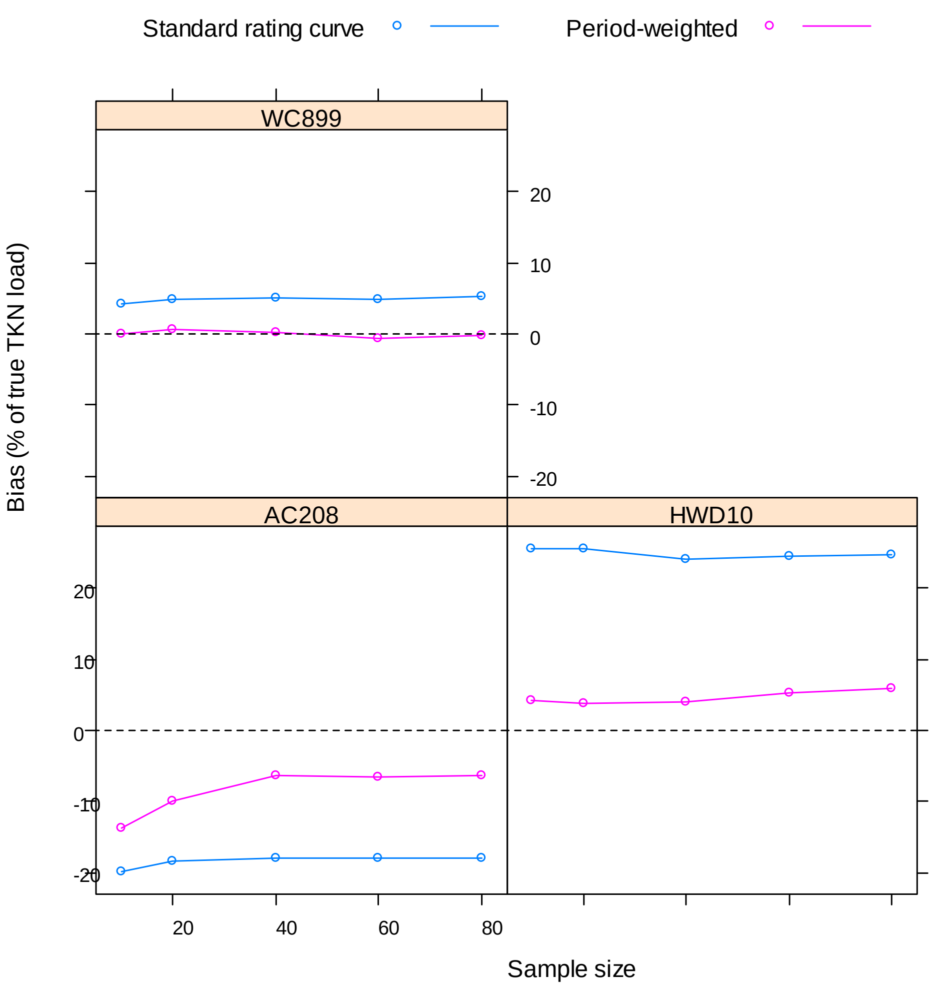


Figure 10-1. Bias in TKN load estimates for standard rating curve and period-weighted estimates, from simulations (24-hr sampling).

We also looked at the differences between historic loads computed by standard rating curves and the selected methods. Loads for this analysis were computed from all LTIMP samples, not just daytime samples. Again, we find that rating curves give generally higher estimates than our selected methods (e.g. Figure 10-2 and Appendix A-4.2). For all constituents and most locations, differences are overwhelmingly positive. The only example with mostly negative differences is SRP at UT-5.

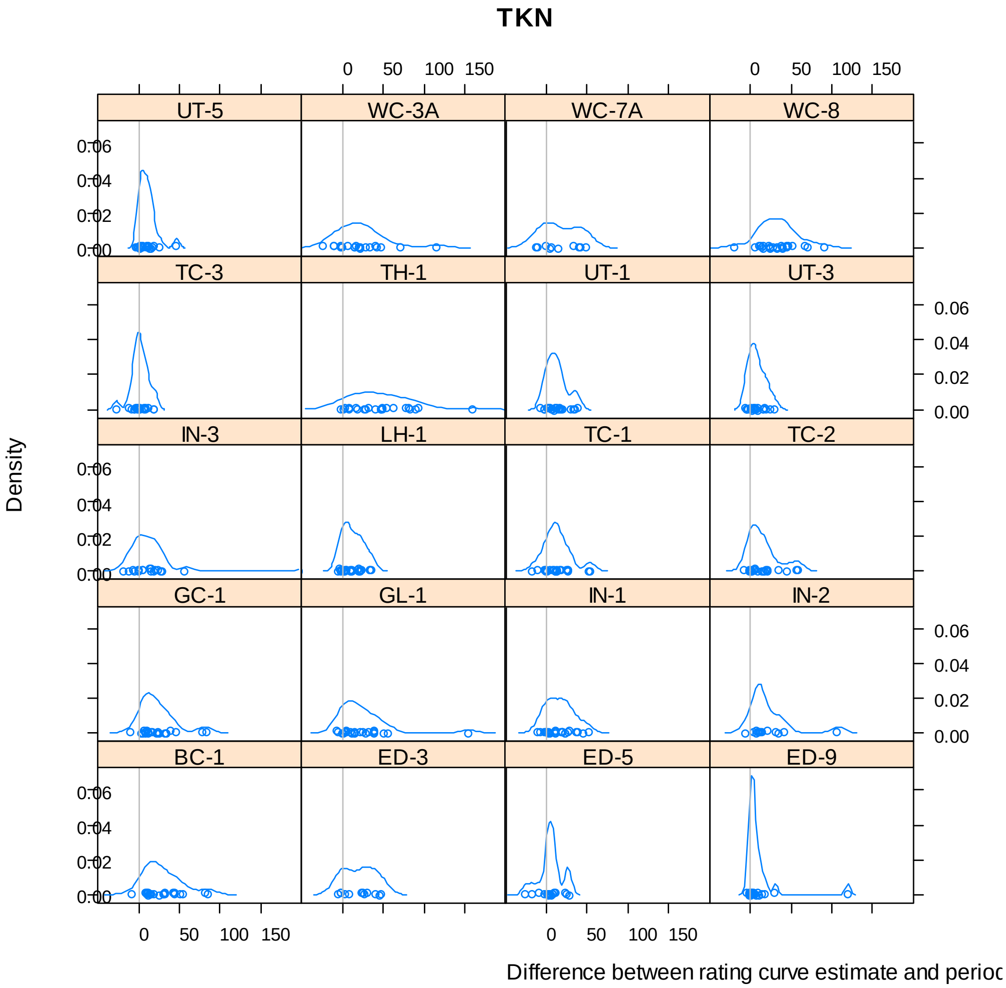


Figure 10-2. Difference between standard rating curve and period-weighted estimates of historic loads for TKN.

It is instructive to compare the sampling requirements for load calculations by the standard rating curve method and the methods developed in this study. Table 10-1 shows the comparison. For particulate constituents (SSC and TP), there are modest gains in efficiency (expressed as a reduction in the number of required samples. For the dissolved constituents, the gains are large. For nitrate-N, for example, using the period-weighted sample method instead of the simple rating curve method reduces the required sample number (at the 90/20) level from 74 to 19 samples.

Table 10-1. Number of samples required to achieve a given level of confidence in load estimates by the standard rating curve method compared with new recommended methods.

		Confidence Level, Pct	80				90				95			
		Error band, as Pct Est. Load	50	30	20	10	50	30	20	10	50	30	20	10
Constituent	**Estimation Method**		**Required Number of Samples per year**											
SS	Simple rating curve		<8	18	50	>100	9	32	85	>100	15	41	>100	>100
Nitrate-N	Simple rating curve		<8	16	57	>100	11	45	74	>100	18	59	>100	>100
TKN	Simple rating curve		<8	14	>100	>100	11	29	>100	>100	<8	36	>100	>100
SRP	Simple rating curve		<8	<8	13	69	<8	15	52	91	<8	22	72	86
Total P	Simple rating curve		<8	<8	26	>100	<8	14	38	>100	<8	20	53	>100
SS	Best model by GRMSE		<8	15	36	>100	<8	27	67	>100	13	36	87	>100
Nitrate-N	Period-weighted		<8	<8	14	33	<8	11	19	43	<8	16	29	51
TKN	Period-weighted		<8	<8	15	60	<8	9	19	75	<8	12	28	85
SRP	Period-weighted		<8	<8	<8	31	<8	8	17	43	<8	15	21	53
Total P	Best model by GRMSE		<8	<8	16	>100	<8	10	25	>100	<8	13	34	>100

11 Time trend analysis on Total Load residuals; Hypotheses to explain observed trends

To evaluate historic trends, we developed regression models for constituent loads to explain natural variability from variations in weather characterized by annual maximum daily flow (*max*) and annual runoff volume (*flow*). Multiple regression models were formulated to include all gaging stations. Location was treated as a categorical variable. Interactions between location and annual peaks and flows were included in these models, when significant, to permit variation in coefficients between watersheds. The model with all potential terms is represented as follows:

$$\log(load) = b_{0j}\, stn_j + b_1 \log(flow) + b_2 \log(max) + b_{3j}\, stn_j{*}\log(flow) + b_{4j}\, stn_j{*}\log(max)$$

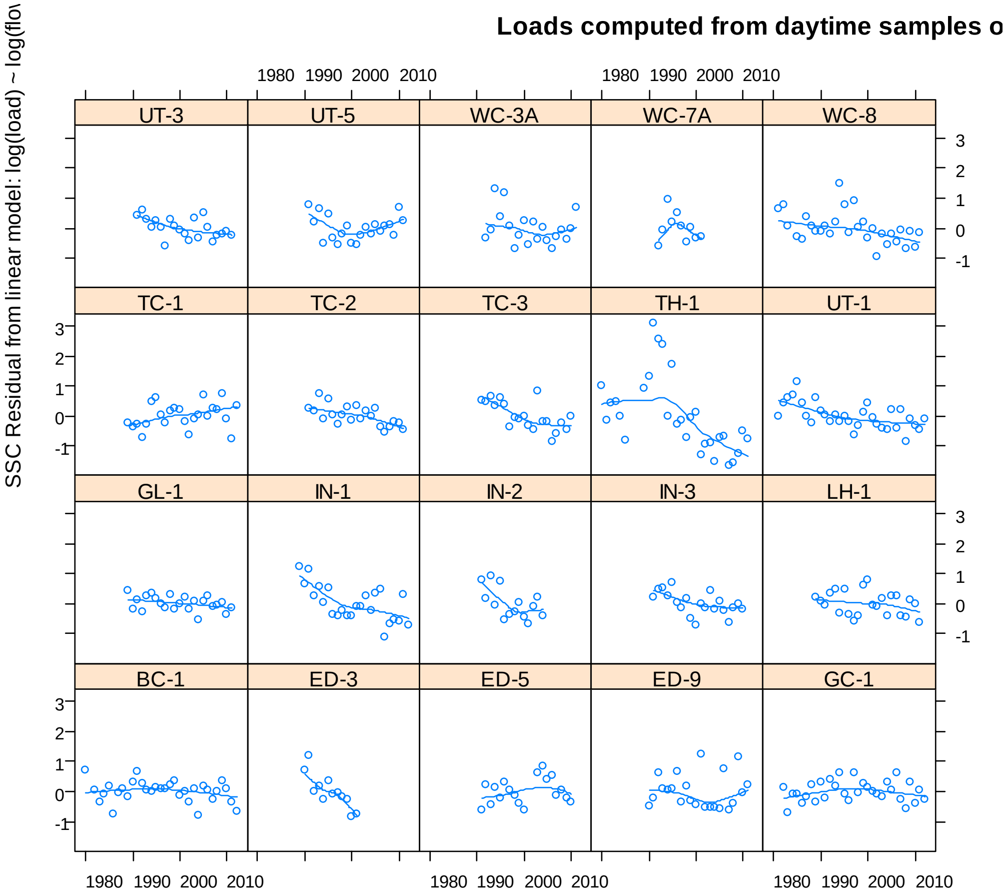
In the best model for each constituent one or more of these terms are zero (Table 11-1). The residuals from each of these models were tested by station for monotonic trend using the "adjusted variable" Mann-Kendall test (Alley, 1988) recommended by Helsel and Hirsch (2002). Alley's test is basically a Mann-Kendall test on the partial regression plot of log(load) versus water year. In the partial regression plot, log(load) and water year are both regressed on the same set of predictors (i.e. those in Table 11-1) and the two sets of residuals are plotted against one another and tested for monotonic trend using the usual Mann-Kendall test. Since there were 20 stations and 20 trend tests, the family-wise Type I error rate was kept to 0.05 using the Bonferroni correction, i.e. the critical p-value was set to *alpha*=0.05/20=0.0025 for each test.

Table 11-1. Models for constituent loads to account for hydrologic variability

	Model	R^2	
		all	day
SS	$\log(\mathit{load}) = b_{0j}\,\mathrm{stn}_j + b_1 \log(\mathit{flow}) + b_2 \log(\mathit{max}) + b_{4j}\,\mathrm{stn}_j{*}\log(\mathit{max})$	.894	.897
TP	$\log(\mathit{load}) = b_{0j}\,\mathrm{stn}_j + b_1 \log(\mathit{flow}) + b_2 \log(\mathit{max}) + b_{3j}\,\mathrm{stn}_j{*}\log(\mathit{flow})$	.921	.928
TKN	$\log(\mathit{load}) = b_{0j}\,\mathrm{stn}_j + b_1 \log(\mathit{flow}) + b_2 \log(\mathit{max})$	.882	.882
NO3	$\log(\mathit{load}) = b_{0j}\,\mathrm{stn}_j + b_1 \log(\mathit{flow}) + b_2 \log(\mathit{max}) + b_{3j}\,\mathrm{stn}_j{*}\log(\mathit{flow})$	.861	.861
SRP	$\log(\mathit{load}) = b_{0j}\,\mathrm{stn}_j + b_1 \log(\mathit{flow}) + b_2 \log(\mathit{max}) + b_{3j}\,\mathrm{stn}_j{*}\log(\mathit{flow})$	.952	.954

11.1 Trend results

If runoff is changing as a result of climate change, there are probably corresponding changes in loading that these analyses would not detect. However, using the same methods as for loads, we found no systematic trends in maximum flow or flow volume. Hence, the partial regression plots (Appendix A-5) look very similar to more familiar plots (Figures 11.1-1 to 11.1-5) of residuals against water year.

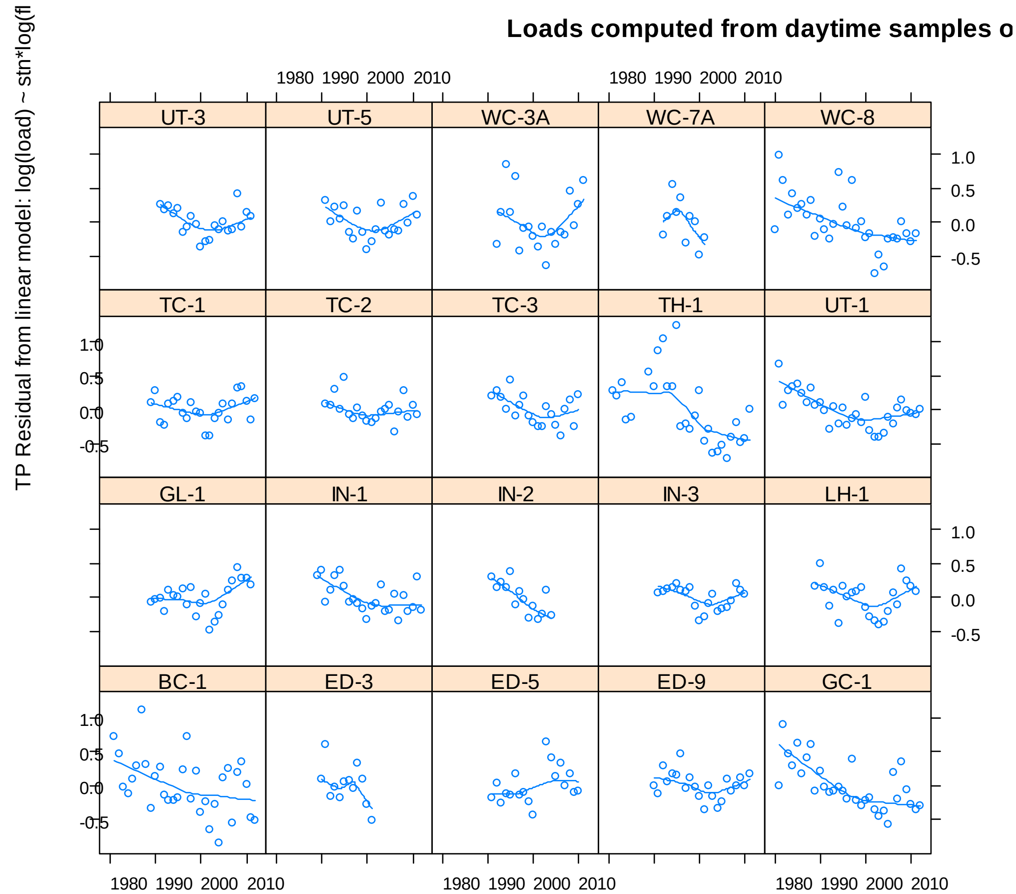


1915

1916 Figure 11.1-1. Trends in SSC after accounting for inter-annual variation in total and maximum daily runoff.

1917

1918



1919

1920 Figure 11.1-2. Trends in TP after accounting for inter-annual variation in total and maximum daily runoff.

1921

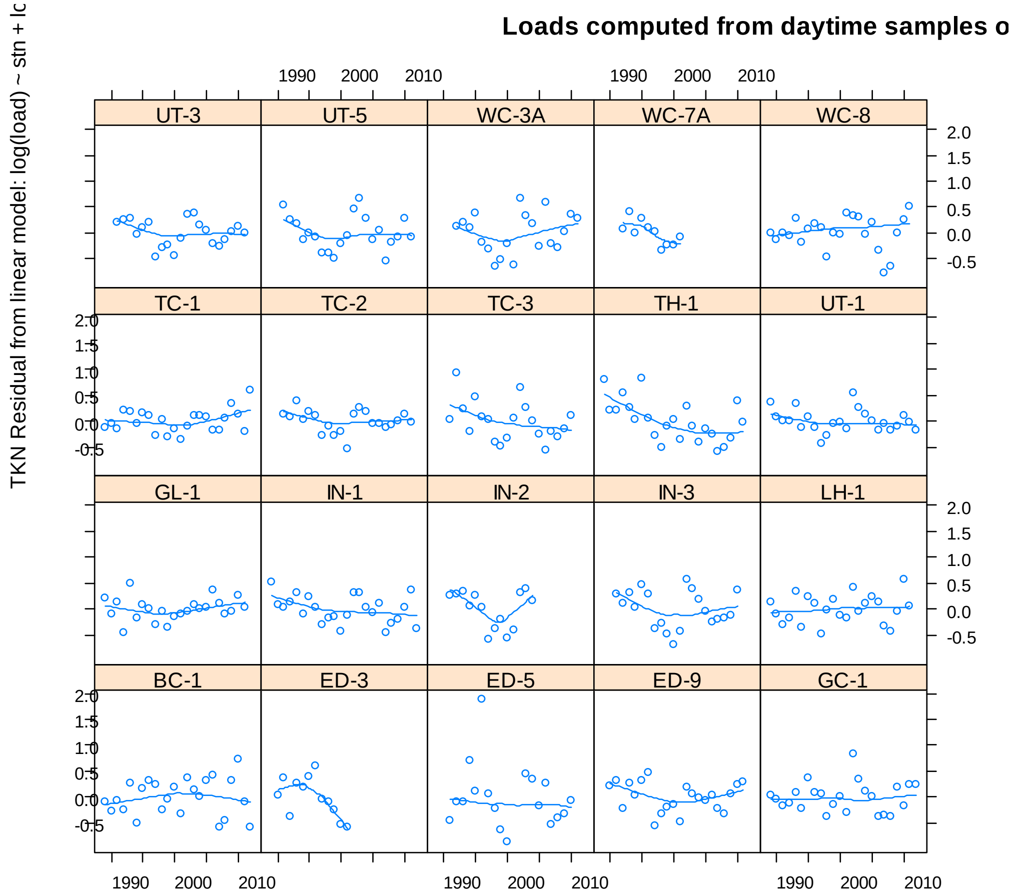


Figure 11.1-3. Trends in TKN after accounting for inter-annual variation in total and maximum daily runoff.

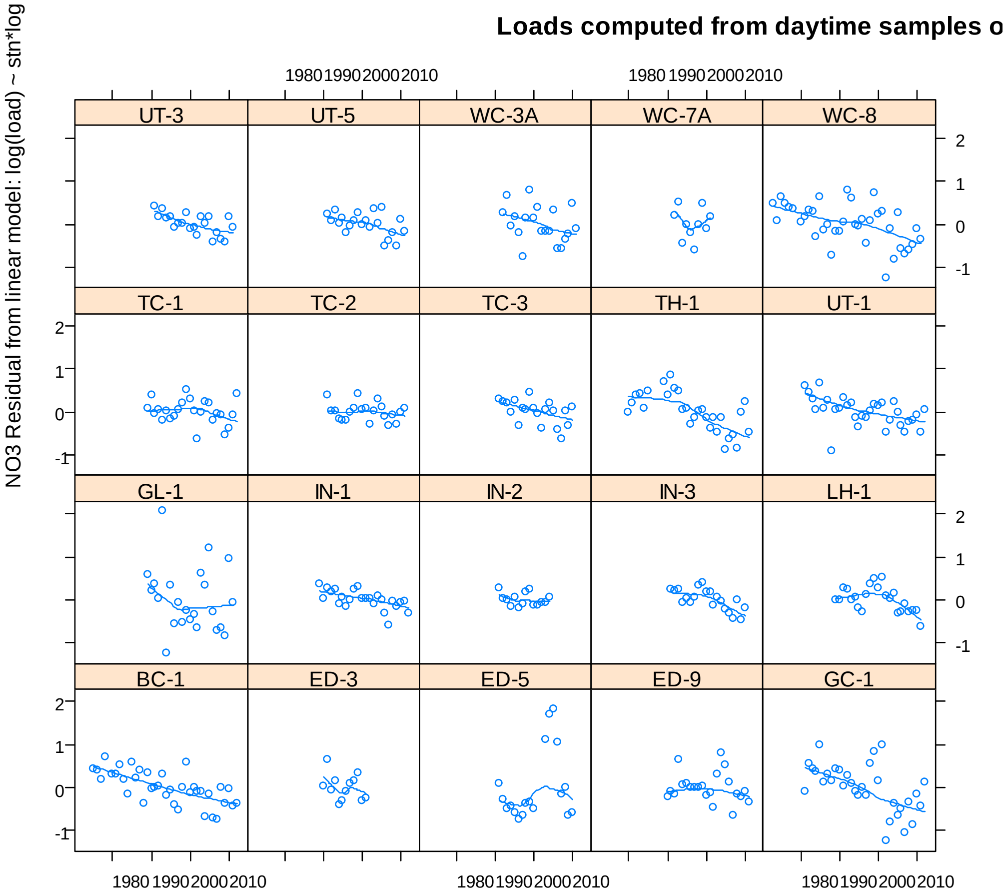


Figure 11.1-4. Trends in NO3 after accounting for inter-annual variation in total and maximum daily runoff.

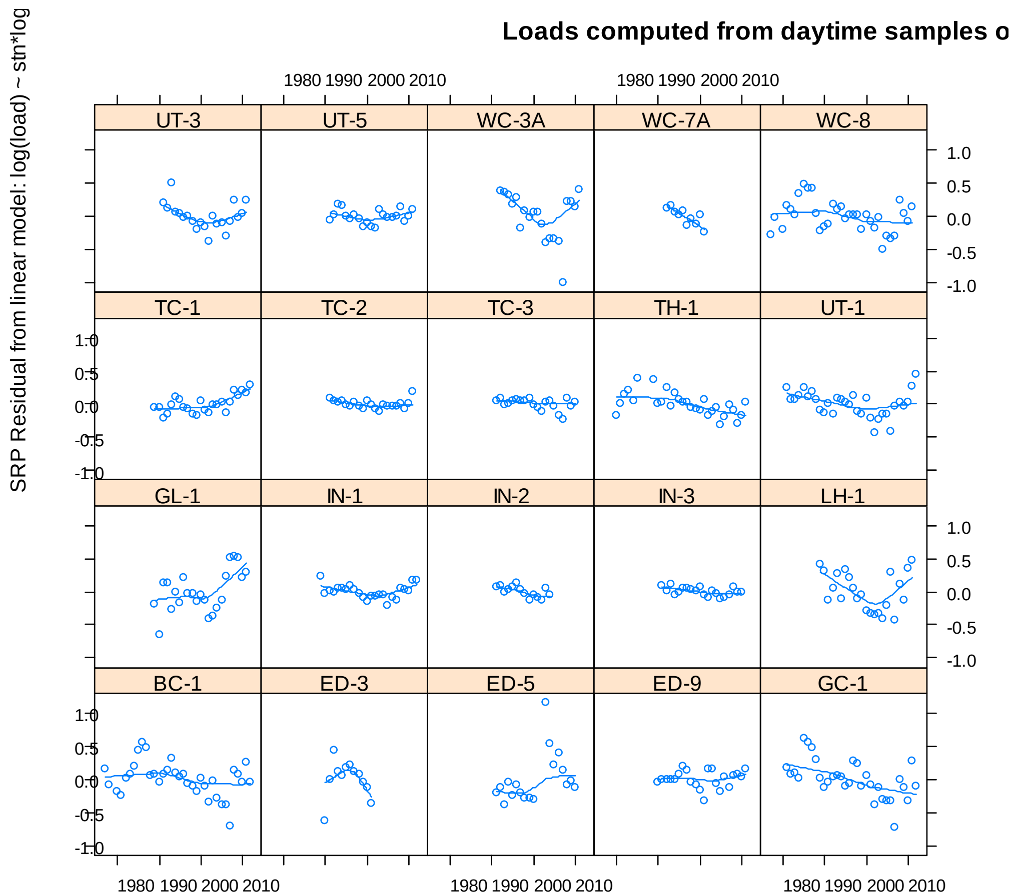


Figure 11.1-5. Trends in SRP after accounting for inter-annual variation in total and maximum daily runoff.

We have included the trend plots for loads based only on daytime samples. Tables in Appendix A-5 show p-values for loads based on all samples as well as those based on daytime samples. The significant trends for loads based on daytime samples were as follows:

- **SS**: ED-3, IN-1, TH-1, UT-1
- **TP**: GC-1, IN-2, TH-1, UT-1, WC-8
- **TKN**: ED-3
- **NO_3-N**: BC-1, GC-1, IN-1, IN-3, TH-1, UT-1, WC-8

- **SRP**: GC-1, TC-1, TH-1

All significant trends are downwards, with the exception of SRP at TC-1. The significant trends at IN-2 and ED-3 are for short periods of record. Those sites have not been sampled since 2006 and 2001, respectively.

Apparent trends in Third Creek (TH-!) could be due to a restoration project. In summer of 2004, the confluence of Rose Creek with Third Creek was relocated from just below Highway 28 to its historic location just north of Lakeshore Blvd. (Susfalk et al., 2010). With the re-diversion, the flow of Rosewood Creek now bypasses the gage and LTIMP sampling site. According to Susfalk (2010), however, the runoff of Rosewood Creek is only about 10 percent of the runoff of Third Creek, and most of its sediment load originates during snowmelt in the upper part of the watershed. A plot of annual runoff at TH-1 vs. annual precipitation at Tahoe City showed no change in the relationship after the rediversion. Neverthless, we tested for time trends in residuals at TH-1 for the period 1981-2004 (day-only samples), and found significant downward trends for SSC, TP, TKN. NO3-N and SRP ($P < 0.02$ for all constituents. We conclude that the downward trends in loads at TH-1 are not an artifact of the re-diversion of Rosewood Creek.

Visually, the TP trends for BC-1 and IN-l look compelling but they don't pass at $p<=0.0025$. A few stations have a concave upward pattern including UT-1 and GC-1, which seem to no longer be declining. TH-1 has the lowest p-value but the first and last 5 years don't fit the trend. A majority of stations have a concave upwards pattern for TP, declining before the year 2000 and flattening or increasing in recent years. This pattern is also evident in a plot of the pooled residuals (Figure 11.1-6).

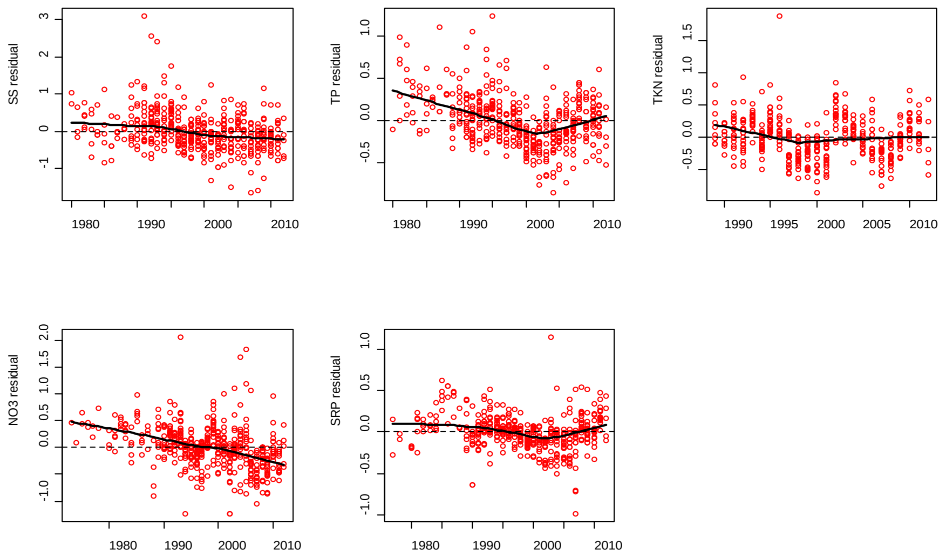


Figure 11.1-6. Trends in SS, TP, TKN, NO3, and SRP after accounting for inter-annual variation in total and maximum daily runoff. Residuals are pooled from all gaging stations shown in Figures 10-1 to 10-5.

There are no compelling trends in TKN. The only significant trend is for the short 12-year record at ED-3.

Most stations have a decreasing pattern of NO_3-N loads, but only 7 of the trends are statistically significant at $p<=0.0025$. The trend for the pooled residuals (Figure 11.1-6) is linear, highly significant ($p< 2e-16$), and represents an average decrease of 2% per year since the 1970s or a total reduction of about 57%. However, small and large watersheds are equally weighted in this calculation.

While SRP had significant trends at 3 locations, one was upwards, and none are very steep. The overall pattern however starting in 1985 is similar to that of TP, with declining loads reversing around 2003.

The occurrence of so many downward trends in loads, especially for NO_3-N, is striking. We hypothesize that the trends are caused by long-term recovery from logging and overgrazing in the 19th century and first half of the 20th century. Essentially, the forests are accumulating biomass, and becoming more effective in retaining nitrogen, phosphorus and sediment. In the case of Blackwood Canyon, recovery may involve a shift from nitrogen-fixing alder toward conifers, which produce a litter and humus layer with high carbon-nitrogen ratio (Coats et al., 1976). The long-term trend toward warmer temperatures could accelerate plant growth and contribute to closing of nutrient cycles and reductions in sediment production.

12 Statistical power analysis

Error in the estimate of a load is a type of measurement error, and it adds error to the regression model that is used in trend detection. Appendix A-6 shows specifically how this error propagates. Because measurement error inflates the regression error it erodes the statistical power to detect trends, and it will take longer to detect a trend than it would with perfect measurements. Interestingly, negative bias increases the regression variance while positive bias decreases it. This confuses interpretations, because methods like the simple rating curve often have positive bias, which makes them look better than methods with zero bias in a statistical power analysis.

12.1 Power analysis methodology

To estimate the power of the Mann-Kendall test, for sample sizes of 10 to 80 years, we randomly resampled the residuals that were tested in the adjusted Mann-Kendall tests for trend. All stations are pooled in these models. The reason for using a resampling procedure rather than generating normally distributed errors is that the residuals from the models, except those for TKN, were somewhat long in the tails (Figure 12.1-1). We also generated normally distributed measurement errors based on specified values of relative error (0.1, 0.3, or 0.5) and bias (-0.2 to +0.2). Our first analysis indicated that bias of less than 10% had very little influence on statistical power (Appendix A-6.2). Subsequently only bias factors of -0.2, 0.0, and +0.2 were used. Before adding the specified measurement error, the model residuals were shrunk by a multiplicative factor (<1) designed to approximately remove the original measurement error from the residual variance (Appendix A-6.3). Finally, a trend of a specified magnitude was added (1, 3, or 5% per year), and the Mann-Kendall test was applied to determine whether the added trend was detectable at alpha levels of 0.0025, 0.0050, and 0.05. The first two significance levels correspond to the Bonferroni rule for testing groups of 20 and 10 stations, respectively. The statistical power is estimated as the proportion of significant tests in 5000 resamplings.

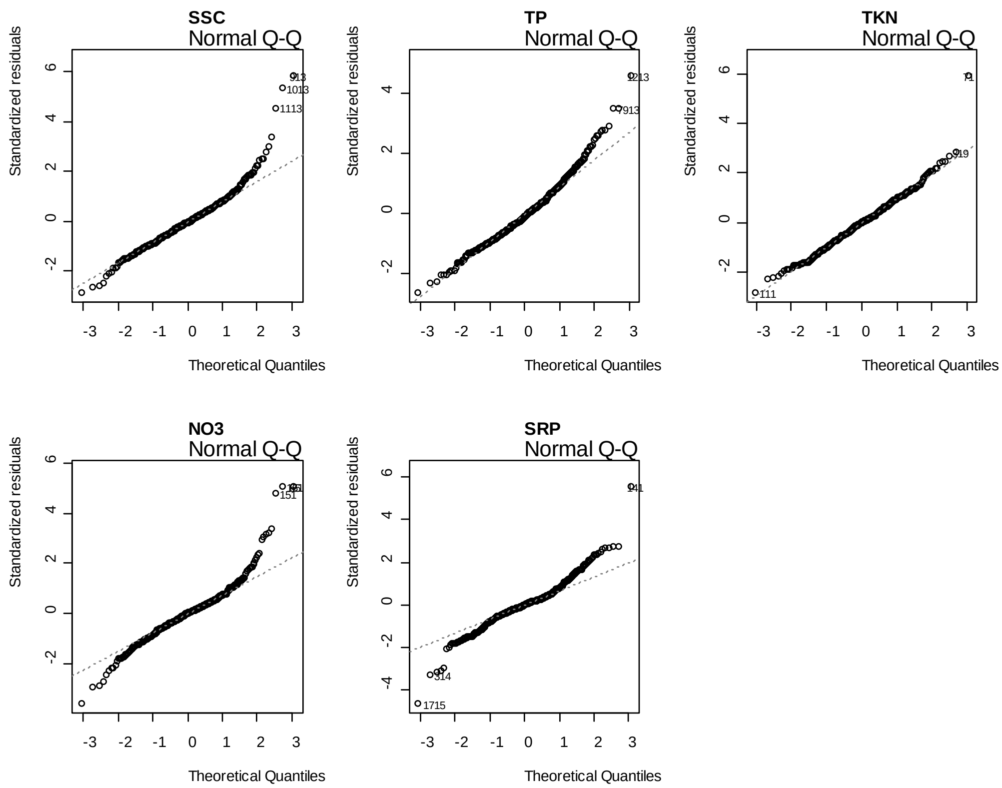


Figure 12.1-1. Quantile-quantile plots of the residuals for models used in the trend tests (Table 10-1). A linear pattern suggests a normal distribution.

12.2 Power analysis results

Figure 12.2-1 shows the power analysis for TP at the 0.005 significance level. The CV (coefficient of variation) is the measurement error as a proportion of the load. In the central frame of the figure it can be seen that to detect a trend of 3% per year with measurement errors averaging 50%, a sample size of 50 years would be needed to attain the same statistical power (close to 1) as 40 years with measurement errors of 30%, or 30 years with measurement errors of 10%. For a stronger trend (left column), statistical power is very good for sample sizes of 40 years or more, regardless of bias or measurement error. However to detect a 5% per year trend within 20 years with 80% probability requires that measurement error be limited to 10%

(regardless of bias). It is virtually impossible to detect a 5% trend in 10 years, regardless of precision and bias. To detect a more subtle trend (right column) with 80% probability or better, would require at least 50 years of measurements with 10% error. Power analysis results for the other constituents and other significance levels can be found in Appendix A-6.

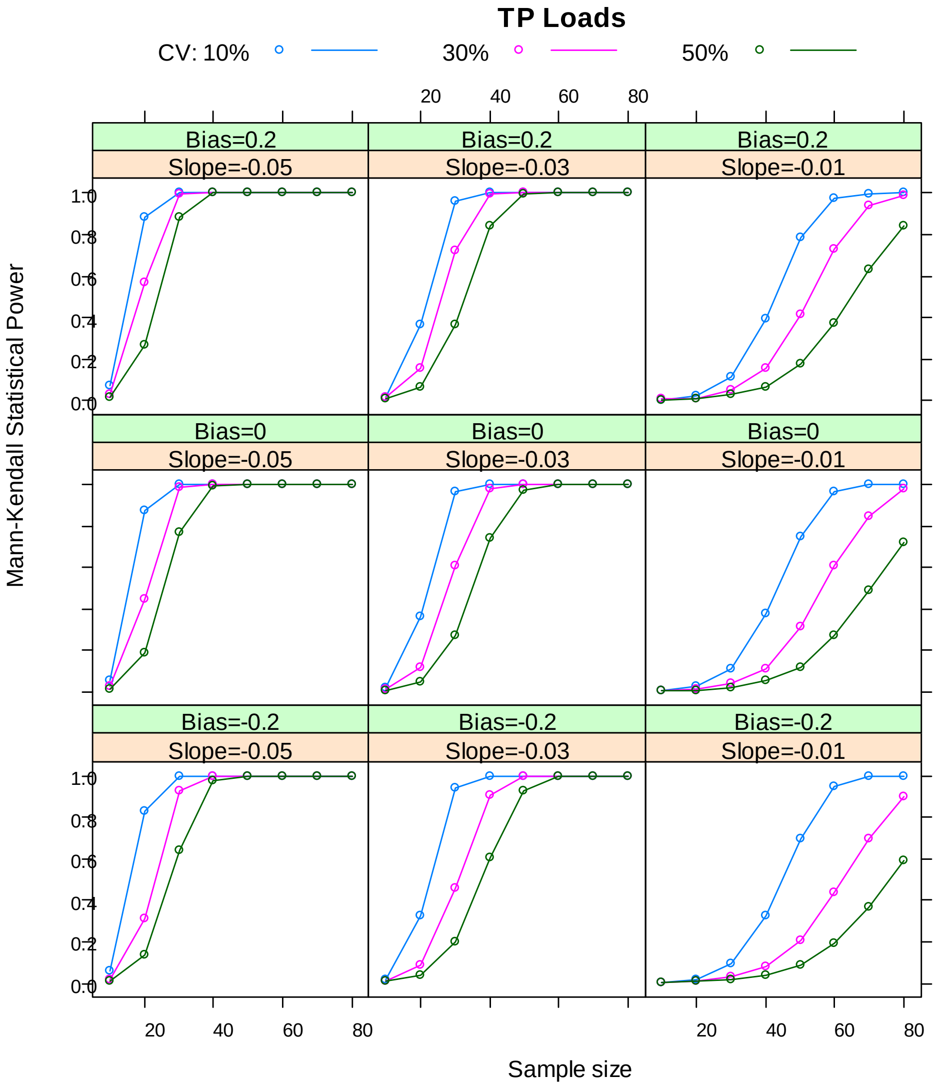


Figure 12.2-1. Power analysis for TP loads, contrasting 3 levels of relative measurement error (CV), for 3 bias

levels and slopes, at 0.005 significance level.

13 Continuous Turbidity Monitoring—A promising technology

In recent years, continuous turbidity monitoring has seen increasing use in the Tahoe basin for estimating sediment and total phosphorus concentrations (see Sec. 4.2). Its application in the LTIMP presents an opportunity to improve the accuracy of total load estimates and to reduce project costs. Most commercially-available probes come with the built-in capacity to measure other parameters, including temperature, pH and conductivity. The USGS has recently been investigating the possibility of retrofitting some existing gaging stations for turbidometry, and has developed cost estimates. The usefulness of continuous turbidity monitoring depends on the relationships between turbidity and the response variable (TP, SSC or FS). Figure 13.1 (from Figure B1, Susfalk et al., 2002) shows a set of graphs of SSC (mg/l) vs. turbidity (NTUs) for a station on Rosewood Creek at State Route 28, for five water years. The upper and lower 95 percent confidence limits on a new prediction are shown in red and blue. Regression equations Note that the data are untransformed. We found some improvement of fit by taking the square root of both variables.

Figure 13.2, from Stubblefield et al. (2007), shows the relationships between TP and turbidity for three years of pooled data from Ward Creek, and one year of data from Blackwood Creek. By including hydrologic variables in the regression, the precision of the estimates TP concentration can be improved.

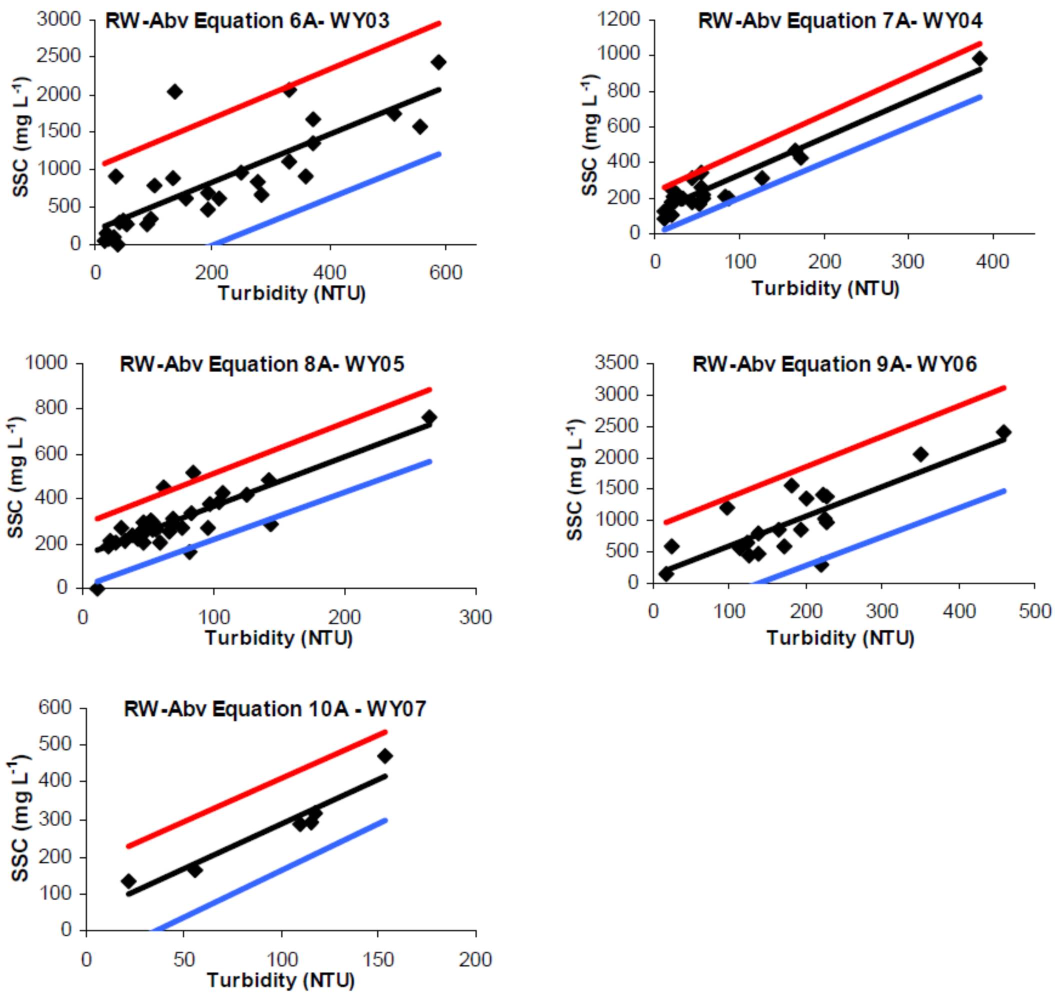


Figure 13.1 SSC (mg/l) vs. turbidity (NTUs) for Rosewood Creek at State Route 28, for five water years. Note that each water year's data are plotted at a different scale. From Susfalk et al., 2007

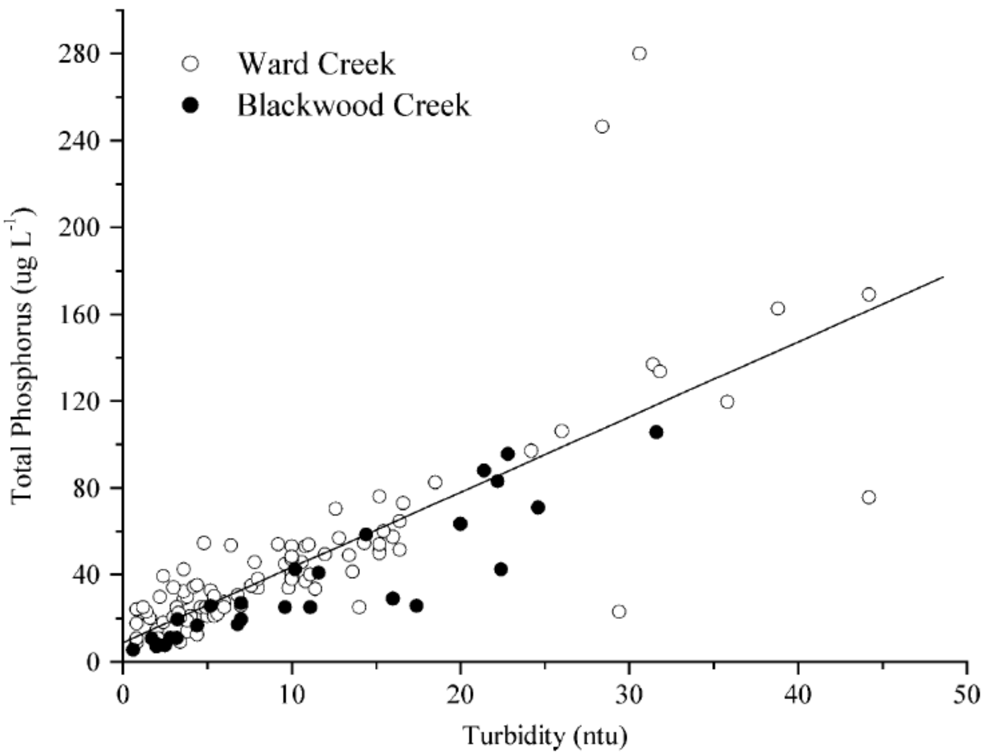


Figure 13-2. TP concentration vs. Turbidity during snowmelt , for Ward Creek (pooled data, 1999, 2000 and 2001, and Blackwood Creek, 2001. From Stubblefield et al., 2007

14 Sample Size, Accuracy and Confidence Limits

In the simulations we tracked the 50th, 80th, 90th, and 95th percentiles of the absolute value of error. These percentiles represent empirical confidence levels for the error. From this information we plotted percent errors for each percentile (confidence level) as a function of sample size for each sampled population (e.g. Figure 14-1). Curves were averaged across populations (Figure 14-1) and interpolated to find the sample sizes necessary (on average) to produce an estimate of specified precision (10, 20, 30, and 50%) at confidence levels of 80, 90 and 95%. Curves for all constituents are provided in Appendix A-2.3. Sample numbers shown as 8 or 100 represent extrapolations at the low and high ends, and should be read as “<8” or “>100”.

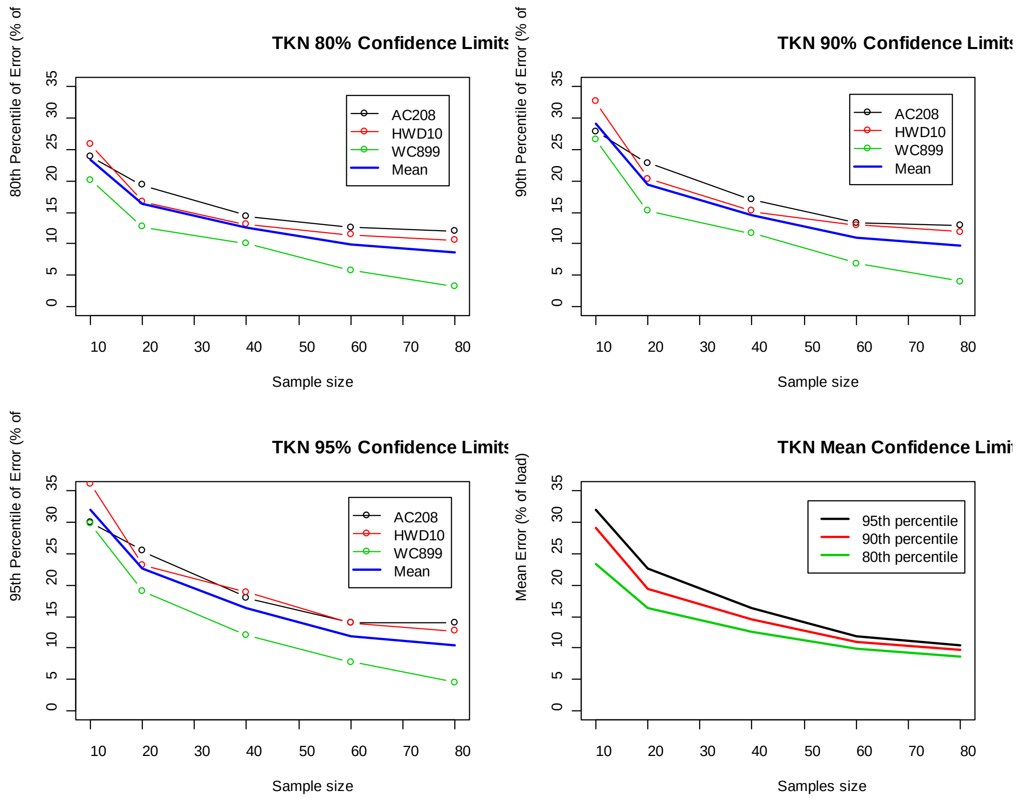


Figure 14-1. Confidence limits on errors for TKN.

14.1 Achieving specified levels of precision in total load

Using the developed relationships between error and sample size for each constituent, we created a spreadsheet showing alternative levels of precision for each constituent as a function of sample size (Table 14-1). The sample numbers refer to the required number of samples per year at a given station. For example, with 25 samples per year, one can be 90 percent sure that the true annual load of total phosphorus is within +/- 20 percent of the value estimated using the best model selected by the GRMSE criterion. For the same level of confidence and percent error for SSC, 67 samples per year would be required. Note that relationship between confidence level and sample number is highly non-linear. To achieve the 90/10 level for TP and SSC would require over 100 samples per year.

Table 14-1. Number of samples required to achieve a given level of precision at a given level of confidence in estimates of total annual load

Confidence Level, Pct		80				90				95			
Error band (+/-), as Pct Est. Load		50	30	20	10	50	30	20	10	50	30	20	10
Constituent	**Estimation Method**	**Required Number of Samples per year, without turbidity**											
SSC	Best model by GRMSE	8	15	36	100	8	27	67	100	13	36	87	100
Fine Sed by mass	Best model by GRMSE	8	15	36	100	8	25	69	100	13	33	83	100
Fine Sed by count	log(FSP) ~ log(q)	8	11	17	60	8	15	26	85	9	19	34	86
Nitrate+nitrite-N	Period-weighted	8	8	14	33	8	11	19	43	8	16	29	51
Ammonium-N	Period-weighted	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
TKN	Period-weighted	8	8	15	60	8	9	19	75	8	12	28	85
SRP	Period-weighted	8	8	8	31	8	8	17	43	8	15	21	53
Total P	Best model by GRMSE	8	8	16	100	8	10	25	100	8	13	34	100
		Required Number of Samples per year with turbidtiy											
SS with turbidity	sqrt(SSC) ~ sqrt(turb)	8	8	13	49	8	8	20	70	8	14	27	96
FS by mass with tu	sqrt(FS) ~ sqrt(turb)	8	8	8	31	8	8	12	47	8	8	16	65
TP with turbidity	Best model by AIC	8	8	9	89	8	8	17	100	8	11	24	100
Lab NTU	Nephelometer	8	8	13	89	8	8	20	100	8	14	27	100
Max no. PWS		8	8	15	60	8	11	19	75	8	16	29	85
Max. Turb-related		8	8	13	89	8	8	20	100	8	14	27	100
Total sample no. w/turb.		8	8	15	89	8	11	20	100	8	16	29	100

Table 14-1 also shows the comparison of required sample numbers for an improved LTIMP (with new load calculation models) with a turbidity-based program, with loads of the particulate constituents calculated by regression with continuous turbidity. For SSC at the 90/20 level, the required sample size drops from 67 to 20.

We obtained estimates of project costs from TERC and USGS for station operation and maintenance and sample collection and analysis. Agency (USGS) overhead costs are included. For the chemical constituents, we multiplied the cost per sample by the maximum number of required samples for the most-variable constituent, assuming that all chemistry samples sent to the lab will be analyzed for all constituents. Aggregating the costs separately for the programs with and without continuous turbidity indicates that a turbidity-based program would be somewhat more expensive than the current program (with improved load calculation), except at the 90/20 level, where it would save over $10,000 per station-year. The spreadsheets with formulas are available on disk.

14.2 Detecting differences of a specified magnitude

We can use our confidence limits to determine sample sizes required to detect a specified relative difference in annual load above and below a project area. Given confidence intervals for two statistics (in our case, loads) Zou and Donner (2008) present a simple method for computing a confidence interval for their difference.

$$L = \hat{\theta}_1 - \hat{\theta}_2 - \sqrt{\left(\hat{\theta}_1 - l_1\right)^2 + \left(u_2 - \hat{\theta}_2\right)^2}$$

$$U = \hat{\theta}_1 - \hat{\theta}_2 + \sqrt{\left(u_1 - \hat{\theta}_1\right)^2 + \left(\hat{\theta}_2 - l_2\right)^2}$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the two estimated loads, (l_1, u_1) and (l_2, u_2) are their confidence intervals and (L,U) is the confidence interval for their difference (at the same confidence level as the original intervals). If the original confidence intervals are symmetric, then the confidence interval for the difference will be symmetric around the difference in estimated loads. Thus a reasonable test of the hypothesis of no difference can be specified by the inclusion or exclusion of zero in the confidence interval for the difference.

Let's assume that the difference in TP loading above and below a project is 30% of the upstream load. Confidence intervals in percentage of the upstream load~~,~~ are (81.9, 118.1) for the upper location and (106.5, 153.5) for the lower location (see Figure A-2-xx, Appendix A-2). With sample sizes of 40, the confidence interval for the difference is (0.4, 59.6), just small enough to exclude zero. Thus a sample size of 40 is required to detect the 30% difference in TP loading. Figure 14.2-1 shows the sample sizes required to detect a specified difference in loading at the 95% confidence level, for each of the major LTIMP nutrients, based on our simulations. With the benefit of continuous turbidity measurements, required sample sizes are reduced somewhat for TP and drastically for SSC (Figure 14.2-2). Reductions for fine sediment should be similar to those for total suspended sediment.

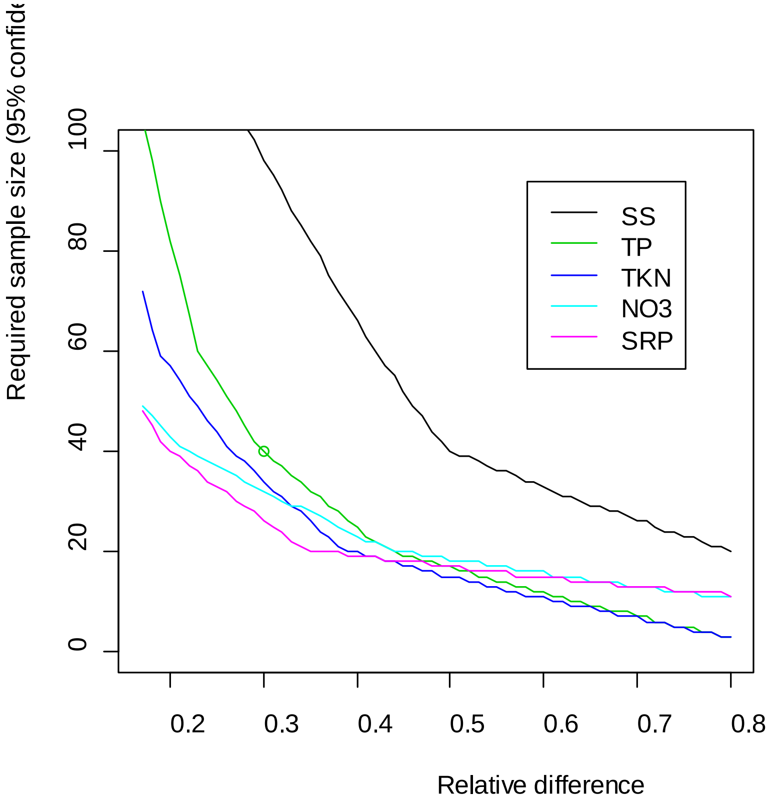


Figure 14.2-1. Sample sizes required to detect a specified difference in loading at the 95% confidence level, for each of the major LTIMP constituents.

Comment [RT5]: Remove the single data point for TP for relative difference of 0.3?

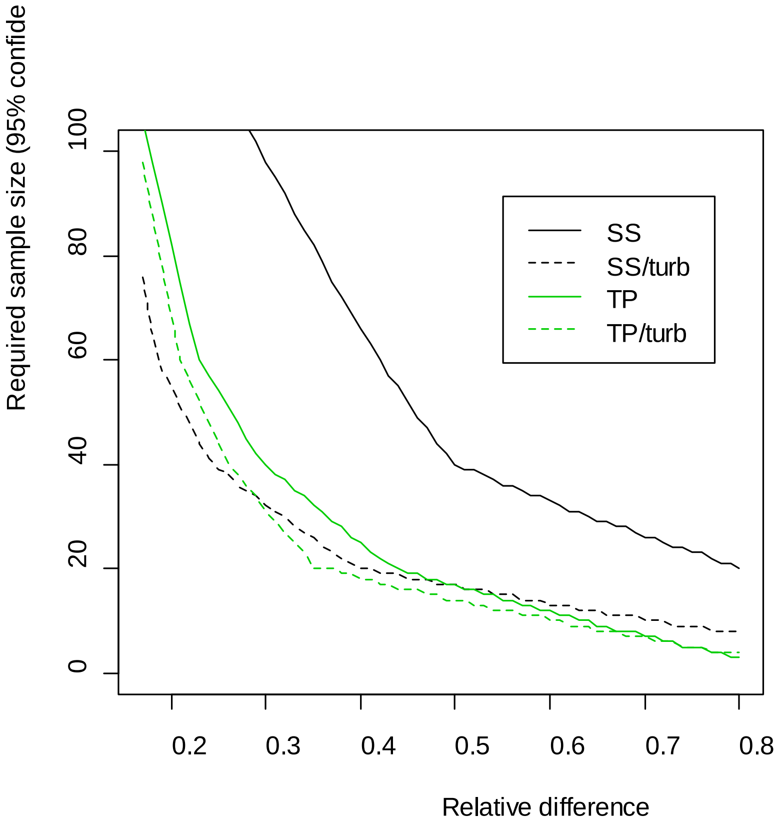


Figure 14.2-2. Sample sizes required to detect a specified difference in loading at the 95% confidence level, for SS and TP, with and without the benefit of continuous turbidity measurements.

15 Summary and Recommendations

In this study, we have developed and compared different methods of calculating total constituent loads, and expressed the results as the number of samples (per station-year) required to achieve a given level of confidence that the true load is within a given error band around the estimated load. Using the best methods (that is, the methods that maximize precision and minimize bias), we recalculated the total annual loads of NO3-N, NH4-N, TKN, SRP, TP and SSC for all of the LTIMP stations over the periods of record. We then related the annual loads to annual runoff and maximum daily peak discharge, and analyzed time trends in the residuals. The significant downward trends indicate a long-term improvement in water quality, which we suggest may be due to long-term recovery of terrestrial ecosystems from 19th and 20th century disturbance.

Based on our results and our experience working with the LTIMP data, we recommend the following:

- Near-continuous measurement of turbidity and temperature with automated probes should become a central part of the realigned LTIMP. Good stream temperature data are important for the Lake Clarity Model, and will be of increasing interest as the climate warms. Turbidity, though by itself not a very important water quality parameter, could play an important role in regression-based estimates of fine sediment and total phosphorus loads. Its use would address the problem of bias in load estimates of TP and SSC resulting from the lack of nighttime samples at some stations. Though at-a-station costs may be somewhat higher with the use of probes, at an intermediate confidence and error-band level (90/20 percent), total costs are actually reduced. Use of the probes should be phased in, with one or two gaging stations equipped in the first year, and more probes added in subsequent years. Turbidity should also be measured in the laboratory in splits of samples collected for fine sediment.

- Fine sediment is now recognized as an important factor in lake clarity and is incorporated in the TMDL targets and used in the Lake Clarity Model. Increased emphasis should be placed on its measurement as number of particles. Additional statistical research on fine sediment is needed in order to refine the error estimates for this constituent, and more work is needed to find the best methods for measuring FS at high concentrations in basin streams. Some of the costs of increased emphasis on fine sediment might be borne by reducing the emphasis on SSC, which is very costly due in part to its high variance, but does not contribute directly to either lake clarity reduction or primary productivity. It may still, however, be of interest in monitoring specific impacts of restoration and mitigation projects.

- The load calculation models developed in this study present an opportunity for major cost savings (or improvements in accuracy of load estimates) especially for the dissolved constituents. These models, written in the programming language R, should be used to update the load estimates for 2013 and 2014, and then used in subsequent years to keep the data base current.

- If station numbers need to be reduced for budget reasons, stations on the big contributing streams (Ward, Blackwood, Trout Creeks and the UTR) should have priority for continued discharge measurement and sampling.

- An intermediate confidence and error level—90 percent confidence that the true value lies within +/- 20% of the estimated value—is achievable with 20 samples per year combined with continuous turbidity, for all constituents, and 25 samples per year without turbidity, for the chemical constituents. Without good turbidity data, this level for SSC would require about 70 samples per year. This intermediate level is perhaps the most realistic one for the program to adopt, since the marginal cost in sample size increases sharply above this level. Managers and planners, however, should be aware of the limitations of the adopted confidence and error level.

- The relationships between sample size, confidence and error should be used to plan monitoring of restoration and mitigation projects.
- Ammonium-N could be dropped from the list of constituents measured, since about half the time its concentration is below the MDL. It is included in the measurement of TKN.
- The LTIMP needs a real director, with a strong background in hydrology, biogeochemistry and statistics along with experience in fund-raising. The director should be housed in one of the scientific research organizations active in the basin (TERC, USGS, DRI/UNR or PSW) and have decision-making authority on operational matters. Overall policy direction would be the responsibility of the management agencies. With continued involvement of support staff from the scientific organizations, this could perhaps be a half-time job. If responsibilities include the urban runoff, air quality and lake monitoring programs, it would be at least a full-time job.

16 References

Allander, K. K., 2004. The effect of a large uncontrolled wildfire on stream nutrient concentration within an undisturbed watershed in the Lake Tahoe Basin [abs.]. *In*: *Research as a Tool in Tahoe Basin Issues, 2nd biennial conference on Tahoe environmental concerns.* Crystal Bay, Nevada, pp. 36-.

Allander, K. K., 2007. Update on the effect of a large uncontrolled wildfire on stream nutrient concentrations within an undisturbed watershed in the Lake Tahoe Basin. *J. Nev. Water Resour. Assoc.*:21.

Alley, W. M., 1988. Using exogenous variables in testing for monotonic trends in hydrologic hime series. *Water Resources Research* **24**:1955-1961.

Alvarez, N. L., A. Pressler, S. Hackley and P. Arneson, 2007. Lake Tahoe Interagency Monitoring Program--An integral part of science in the Lake Tahoe Basin. 2006 Lake Tahoe Special Edition.*Jour. Nev. Water Resour. Assoc.* p. 38

Anderson, C., 2005. Turbidity 6.7 Ver. 2.1. *In*: U.S. Geological Survey Techniques of Water Resources Investigations Book 9. Reston, VA, p. 55.

Lahontan Regional Water Quality Control Board. 1995. Water Quality Standards and Control Measures for the Tahoe Basin. *In*: *Water Quality Control Plan for the Lahontan Region.* State of California, South Lake Tahoe, CA.

Burnham, K. P. and D. R. Anderson, 2002. *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach*. New York, USA, Springer-Verlag.

Burns, D. H., 1990. Real-time sampling strategies for estimating nutrient loadings. *Journal of Water Resources Planning and Management* **116**:727-741.

Carlsson, P., E. Granell and A. Z. Segatto, 1999. Cycling of biologically available nitrogen in riverine humic substances between marine bacteria, a heteroptrophic nanoflagellate and a phytosnthetic dinoflagellate. *Aquat. Microb. Ecol.* **18**:23-36.

Chang, C. C. Y., J. S. Kuwabara, and S. P Pasilis, 1992. Phosphate and iron limitation of phytoplankton biomass in Lake Tahoe. *Can. Jour. Fish. and Aquatic Sci.* **49**:1206-1215.

Clarke, R. T., 1990. Statistical characteristics of some estimators of sediment and nutrient loadings. *Water Resour. Res.* **26**:2229-2233.

Cleveland, W. S. and S. J. Devlin, 1988. Locally-weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**:596-610.

Coats, R. N., M. Costal-Cabral, J. Riverson, J. Reuter, G. B. Sahoo and B. Wolfe, 2013. Projected 21st trends in hydroclimatology of the Tahoe basin. *Climatic Change* **116**:51-69.

Coats, R. N. and C. R. Goldman, 2001. Patterns of nitrogen transport in streams of the Lake Tahoe basin, California-Nevada. *Water Resour. Res.* **37**:405-415.

Coats, R. N., M. Larsen, A. H. J. Thomas, M. Luck and J. Reuter, 2008. Nutrient and sediment production, watershed characteristics, and land use in the Tahoe Basin, California-Nevada. *J. Am. Water Resour. Assoc.* **44**:754-770.

Coats, R. N., R. L. Leonard and C. R. Goldman, 1976. Nitrogen uptake and release in a forested watershed, Lake Tahoe Basin, California. *Ecology* **57**:995-1004.

Coats, R. N., F. Liu and C. R. Goldman, 2002. A Monte Carlo test of load calculation methods, Lake Tahoe Basin, California-Nevada. *Jour. Am. Water Resour. Assoc.* **38**:719-730.

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.

Cohn, T. A., D. L. Cauldr, 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 Resour. Res.* **28**:2353-2363.

Cohn, T. A., 1995. Recent advances in statistical methods for the estimation of sediment and nutrient transport in rivers. *Rev. Geophys.* **33 Suppl.**

Dann, M. S., J. A. Lynch, and E. S. Corbett, 1986. Comparison of methods for estimating sulfate export from a forested watershed. *J. Environ. Qual.* **15**:140-145.

Das, T., E. P. Maurer, D. W. Pierce, M. D. Dettinger and D. R. Cayan, 2013. Increases in flood magnitudes in California under waming climates. *J. Hydrology* **501**:101-110.

Eads, R. E. and J. Lewis, 2003. Turbidity Threshold Sampling in Waterhed Research. *In*: *First Interagency Conference on Research in the Watersheds.,* K.G. Renard, S. A. McElroy, W.J. Gburek, H. E. Canfiled and R. L. Scott (eds). USDA Agric. Res. Serv., pp. 567-571.

Edwards, T.K., and G.D. Glysson, 1999. Field Methods for Measurement of Fluvial Sediment: U.S. Geological Survey Techniques of Water Resources Investigations, Book 3, chapter C2, 89 p.10. (available at http://pubs.usgs.gov/twri/twri3-c2/.)

Ferguson, R. I., 1986. River loads underestimated by rating curves. *Water Resour. Res.* **22**:74-76.

Froelich, P. N., 1988. Kinetic control of dissolved phosphate in natural rivers and estuaries: A primer on the phosphate buffer mechanism. *Limnol. Oceanogr.* **33**:649-668.

Gilroy, E. J., R. M. Hirsch and T. A. Cohn, 1990. Mean square error of regression-based constituent transport estimates. *Water Resources Research* **26**:2069-2077.

Goldman, C. R., 1981. Lake Tahoe: Two decades of change in a nitrogen deficient oligotrophic lake. *Verh. Internat. Verein. Limnol.* **21**:45-70.

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.

Goldman, C. R. and R. Armstrong, 1969. Primary productivity studies in Lake Tahoe, California. *Internationale Vereinigung fuer Theoretische und Angewandte Limnologie Verhandlungen* **17**:49-71.

Goldman, C. R. and R. Armstrong, 1969. Primary Productivity Studies in Lake Tahoe, California. *Verh Internat Verein Limnol* **17**:49-71.

Grismer, M. E., 2013. Stream sediment and nutrient loads in the Tahoe Basin—estimated vs monitored loads for TMDL “crediting”. *Environ Monit Assess* **185**:7883–7894.

Harrell, F. E. et al. 2010. Hmisc: Harrell miscellaneous, R package version 3.8-3.

Hatch, L. K., 1997. The generation, transport, and fate of phosphorus in the Lake Tahoe ecosystem. Univ. of Calif., Davis, p. 212.

Hatch, L. K., J. Reuter and C. R. Goldman, 1999. Relative importance of stream-borne particulate and dissolved phosphorus fraction to Lake Tahoe phyytoplankton. *Can. J. Fish Aquat Sci.* **56**:2331-2330.

Helsel, D. R. and R. M. Hirsch, 2002. Statistical methods in water resources Techniques of Water Resources Investigations. U.S. Geological Survey, p. 522.

Heyvaert, A., D. M. Nover, T. G. Caldwell, W. B. Trowbridge, S. G. Schladow and J. E. Reuter, 2011. Assessment of Particle Size Analysis in the Lake Tahoe Basin. Desert Res. Instit. and Univ. of Calilf. Davis.

Hill, A. R., 1986. Stream nitrate loads in relation to variations in annual and seasonal runoff regimes. *Water Resour. Bull.* **22**:829-839.

Hirsch, R. M., J. R. Slack, and R. A. Smith, 1982. Techniques of trend analysis for monthly water quality data. *Water Resour. Res.* **18**:107-121.

Hirsch, R. M., and J. R. Slack, 1984. A nonparametric trend test for seasonal data with serial dependence. *Water Resour. Res.* **20**:727-732.

Hirsch, R. M., R. B. Alexander, and R. A. Smith, 1991. Selection of methods for the detection and estimation of trends in water quality. *Water Resour. Res.* **27**:803-813.

Jassby, A. D., J. E. Reuter, R. P. Axler, C. R. Goldman, and S. H. Hackley, 1994. Atmospheric deposition of nitrogen and phosphorus in the annual nutrient load of Lake Tahoe (California-Nevada). *Water Resour. Res.* **30**:2207-2216.

Johnson, D. W., W. W. Miller, E. Carroll and W. R.F., 2007. The Gondola fire: Effects on Nutrient Loss, Soil Fertility and Leaching. *Jour. Nev. Water Resour. Assoc.* **38**.

Johnson, K. S. and L. J. Coletti, 2002. In situ ultravioled spectrophtometry for high resolution and long-term monitoring of nitrate, bromide and bisulfide in the ocean. *Deep-Sea Res. I* **48**:1291-1305.

Kamphake, L. J., S. A. Hannah and J. M. Cohen, 1967. Automated analysis for nitrate by hydrazine reduction. *Water Research* **1**:205-219.

Kempers, A. J., and G. Van Der Velde, 1992. Determination of nitrate in eutrophic coastal seawater by reduction to nitrite with hydrazine. *Jour. Environ. Anal. Chem.* **47**:1-6.

Kempers, A. J. a. L., A. G., 1988. Re-examination of the Determination of Environmental Nitrate as Nitrite by Reduction with Hydrazine. *Analyst* **113**:3-6.

Leonard, R. L., L. A. Kaplan, J. F. Elder, R. N. Coats and C. R. Goldman, 1979. Nutrient Transport in Surface Runoff from a Subalpine Watershed, Lake Tahoe Basin, California. *Ecological Monographs* **49**:281-310.

Lewis, J., 1996. Turbidity-controlled suspended sediment sampling for runoff-event load estimation. *Water Resour. Res.* **32**:2299-2310.

Lewis, J. and R. Eads, 2009. Implementation Guide for Turbidity Threshold Sampling: Principles, Procedures, and Analysis. *In: Gen. Tech. Rep.*, Editor (Editor)^(Editors). USDA For. Serv. Pacific SW Res. Sta., p. 87.

Liston, A., Ja, e;;. V/, V. Edirveerasingam, B. Allen, S. Hackley, J. Reuter and S. G. Schladow, 2013. *Laboratory Procedure, Field Protocol and Quality Assurance Manual*. Davis, CA, Tahoe Environmental Research Center, University of California, Davis.

Lumley, T. and A. Miller, 2009. Leaps: regression subset selection, R package version 2.9.

Miller, R. J., 1981. *Similtaneous Statistical Inference*. New York, Springer Verlag.

Ngai, K. L. C., 2008. Potential effects of climate change on the invasion of largemouth bass (Micropterus salmoides) in Lake Tahoe, California-Nevada. *In*: *Graduate Dept. of Zoology*. University of Toronto, Toronto, Canada, p. 118.

Nover, D. M., 2012. Fine particles in watersheds: measurement , watershed sources and technologies for removal. *In*: *Civil & Environ. Eng*. University of Calilfornia at Davis, p. 98.

Pellerin, B. A., B. A. Bergamaschi, B. D. Downing, J. F. Saraceno, J. D. Garrett and L. D. Olsen, 2013. Optical Techniques for the Determination of Nitrate in Environmental Waters: Guidelines for Instrument Selection, OPeration, Deployment, Maintenance, Quality Assurance, and Data REproting. *In*: *Techniques and Methods 1-D5*. U.S. Geol. Surv., Reston, VA, p. 37.

Pellerin, B. A., J. F. Saracen, J. B. Shanley, S. S.D., G. R. Aiken, W. M. Wolleim and B. A. Bergamaschi, 2012. Taking the pulse of snowment: in site sensors reveal seasonal, event and diurnal patterns of nitrate and dissolved organic matter variabilit in an upland forest stream. *Biogeochemistry* **108**:183-198.

Preston, S. D., V. J. Bierman, Jr., and S. E. Silliman, 1989. An evaluation of methods for estimation of tributary mass loads. *Water Resour. Res.* **25**:1379-1389.

Protection, N. D. o. E., 2013. Nevada 2008-10 Water Quality Integrated Report with EPA Overlisting. *In*, Editor (Editor)^(Editors). Nevada Division of Environmental Protection, Carson City, NV.

R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rabidoux, A. A., 2005. Spatial and temporal distribution of fine particles and elemental concentrations in suspended sediments in Lake Tahoe streams, California-Nevada, University of California, Davis. M.S. Thesis, *Civil Engineering.* Univ. Calif. Davis, 154 pp.

Reuter, J. E., A. Heyvaert, O. A., A. Parra and R. Susfalk, 2012. Water quality conditions following the 2007 Angora wildfire in the Lake Tahoe basin. Univ. of Calif. Davis and Desert Research Instit., Davis, CA, 75 pp.

.

Riverson, J., R. Coats, M. Costal-Cabral, M. D. Dettinger, J. Reuter, G. B. Sahoo and S. G. Schladow, 2013. Modeling the transport of nutrients and sediment loads into Lake Tahoe under projected climatic changes. *Climatic Change* **116**:35-50.

Robertson, D. M. and E. D. Roerish, 1999. Influence of various water quality sampling strategies on load estimates for small streams. *Water Resour. Res.* **35**:3747-3759.

Rowe, T. G., D. K. Saleh, S. A. Watkins and C. R. Kratzer, 2002. Streamflow and Water-Quality Data for Selected Watersheds in the Lake Tahoe Basin, California and Nevada, Through September 1998. *In: Water-Resources Investigations Rep.*, Editor (Editor)^(Editors). U.S. Geol. Surv. Report 02-4030, Carson City, NV, 118 pp..

Sahoo, G. B., D. M. Nover, J. E. Reuter, A. Heyvaert, J. Riverson and S. G. Schladow, 2013. Nutrient and particle load estimates to Lake Tahoe (CA-NV, USA) for Total Maximum Daily Load establishment. *Science of the Total Environment* **444**:579-590.

Sahoo, G. B., S. G. Schladow, J. Reuter, R. Coats, M. D. Dettinger, J. Riverson, B. Wolfe and M. Costal-Cabral, 2013. The response of Lake Tahoe to climate change. *Climatic Change* **116**:71-95.

Sauer, V. B. and D. P. Turnipseed, 2010. Stage Measurement at Gaging Stations. *In*: *U.S. Geological Survey Techniques and Methods book 3, chap. A7.* p. 45.

Seitzinger, S. P., R. W. Sanders and r. Styles, 20002. Bioavailability of DON from natural and anthropogenic sources to estuarine phytoplankton. *Limnol. Oceanogr.* **47**:353-366.

Simon, A., 2006. Estimates of Fine-Sediment Loadings to Lake Tahoe from Channel and Watershed Sources. USDA-Agric. Res. Serv., National Sedimentation Laboratory, Oxford MS.

Simon, A., E. Langendoen, R. Bingner, R. Wells, A. Heins, N. Jokay and I. Jaramillo, 2003. Draft Final Lake Tahoe Basin Framework Implementation Study: Sediment Loadings and Clhannel Erosion. *In*, Editor (Editor)^(Editors). USDA Agric. Res. Serv. National Sedimentation Laboratory, Oxford MS.

Spector, P., J. Friedman, R. Tibshirani and T. Lumley, 2010. acepack: ace() and avas() for selecting regression transformations, R package version 1.3-3.0.

Stubblefied, A. P., J. E. Reuter, R. A. Dahlgren and C. R. Goldman, 2007. Use of turbidometry to characterize suspended sediment and phosphorus fluxes in the Lake Tahoe basin, California, USA. *Hydrologial Processes* **21**:281-291.

Stubblefield, A. P., 2002. Spatial and Temporal Dynamics of Watershed Sediment Delivery, Lake Tahoe, California. Ph.D. Thesis. University of California Davis, Davis, CA, p. 199.

Susfalk, R., B. Fitzgerald, S. Brown and D. Fellers, 2010. Rosewood Creek Monitoring Compendium 2002-2010. Desert Research Institute and Nevada Tahoe Conservation District, Reno, NV, 141.

Thomas, R. B., 1985. Estimating Total Suspended Sediment Yield with Probability Sampling. *Water Resources Research* **21**:1893-1905.

Thomas, R. B., 1989. Piecewise SALT sampling for estimating suspended sediment yields. *In,* Editor (Editor)^(Editors). Pacific Southwest Forest and Range Exp. Sta., Forest Service, U.D. Dept. Agric. Report PSWA-114, Berkeley, CA, p. 11.

Thomas, R. B. and J. Lewis, 1995. An evaluation of flow-stratified sampling for estimating suspended sediment loads. *Jour. Hydrology* **170**:27-45.

Turnipseed, D. P. and V. B. Sauer, 2010. Discharge Measurements at Gaging Stations. *In*: *U.S. Geological Survey Techniques and Methods book 3, chap. A8.* Reston, VA, p. 87.

Winder, M., Reuter, J.R., and Schladow, S.G., 2008. Lake warming favors small-sized planktonic diatom species. *Proc. Royal Soc. B.* **276**.

Wittmann, M. E., C. Ngai and S. Chandra, 2013. Our New Biological Future? The Influence of Climate Change on the Vulnerability of Lakes to Invasion by Non-Native Species. *In*: *Climatic Change and Global Warming on Inland Waters,* C. R. Goldman, M. Kumagai and R. D. Robarts (eds.) Wiley -Blackwell, West Sussex, UK, pp. 271-294.

Zou, G. Y. and A. Donner, 2008. Construction of confidence limits about effect measures: A general approach. Statistics in Medicine 27:1693-1702.

Acknowledgements

We thank Charles Goldman and Geoffrey Schladow for their many years of leadership of the UC Davis Tahoe Research Group and Tahoe Environmental Research Center; Patty Arneson for careful data stewardship and patiently responding to our questions about the LTIMP data, Nancy Alvarez for helpful suggestions and information about the USGS work in the Basin; Nicole Beck, Mark Grismer, Alan Heyvaert, Andy Subblefied, and Rick Susfalk, for generously sharing their turbidity and discharge data for basin streams; Bob Thomas for helpful suggestions and critical review of the report; Jim Markle for the use of six photographs illustrating LTIMP sampling methods, and Tiff van Huysen for contract administration. This work was funded by grant no. 13-DG-11272170-009 from USDA Forest Service under the Southern Nevada Public Land Management Act.