Trend-oriented sampling strategy and estimation of soluble reactive phosphorus loads in streams

Lorenz Moosmann, Beat Müller, René Gächter, and Alfred Wüest

Limnological Research Center, Swiss Federal Institute for Environmental Science and Technology, Kastanienbaum, Switzerland

Ernst Butscher and Peter Herzog

Umwelt und Energie, Luzern, Switzerland

Received 30 July 2004; revised 9 November 2004; accepted 23 November 2004; published 29 January 2005.

[1] Nutrient transfer from soils to surface waters is associated with large, hydrologically induced fluctuations. Consequently, stream-based estimation of long-term changes in nutrient leaching is masked by variations of stream discharge. Using high-resolution discharge and soluble reactive phosphorus (SRP) data from three small catchments (up to 42 km^2), SRP loads are calculated by applying two different rating curves. Monte Carlo simulations are carried out to determine monitoring strategies for optimizing the number of water samples, their distribution between periods of low and high discharge, and the duration of composite sample collection. Trends in SRP load are isolated from natural variations by applying the discharge time series of 1 year each to annually changing rating curves. By applying this approach to various monitoring data sets, collected over the past 15 years, downward trends in SRP leaching of up to -3% yr^{-1} are detected. We describe how to determine the number of annual samples required to detect trends in nutrient load, depending on monitoring duration, available resources, and the magnitude of the expected trend.

Citation: Moosmann, L., B. Müller, R. Gächter, A. Wüest, E. Butscher, and P. Herzog (2005), Trend-oriented sampling strategy and estimation of soluble reactive phosphorus loads in streams, *Water Resour. Res.*, *41*, W01020, doi:10.1029/2004WR003539.

1. Introduction

[2] Nutrient input to the trophogenic layer is the key factor controlling primary production in lakes, reservoirs and estuaries [*Schindler*, 1978]. For lakes and reservoirs of short to medium hydraulic residence timescales, the largest share of this input is commonly contributed from their catchment areas. Consequently streams and rivers constitute the most important pathway for nutrient input. Owing to high variability in both discharge and nutrient concentrations, the estimation of nutrient loads requires elaborate sampling strategies to ensure resolving the variance of the load [*Richards and Holloway*, 1987].

[3] The most important factors to be taken into account for appropriate resolution of the load are short-term variations in discharge [*Bodo and Unny*, 1983; *Robertson and Roerisch*, 1999], the immediate response of nutrient concentrations to discharge variations [*Vanni et al.*, 2001; *Vieux and Moreda*, 2003; *Gächter et al.*, 2004], and long-term (multiyear or seasonal) changes in discharge or concentration [*Galat*, 1990; *Clement*, 2001]. Owing to high monitoring costs, nutrient concentrations are usually sampled at low frequencies. Various estimation techniques have been developed for determining nutrient loads from sparse monitoring data, such as ratio estimators [*Dolan et al.*, 1981], stratification techniques [*Richards and Holloway*, 1987; *Thomas and Lewis*, 1995], or rating curves [*Cohn et al.*, 1989]. For details of the different methods we refer to the reviews by *Preston et al.* [1989] and *Cohn* [1995].

[4] In recent years, a new water management issue has evolved, that poses additional challenges to sampling strategies, sampling accuracy, and load estimation procedures. In the 1970s and 1980s, wastewater input constituted the main nutrient source in countries like Switzerland. Today, as advanced wastewater treatment is largely in place, the main nutrient sources of many streams shifted from point to diffuse sources, such as leaching from agricultural soils [*Clement*, 2001; *Coats et al.*, 2002]. The implications of this shift are twofold. First, due to the dynamic nature of leaching, nutrient loads show larger variations today than in the past, when wastewater constituted the major input. Second, the reaction of leaching, caused by changes in land use, may be retarded due to large nutrient stocks accumulated in soils. Thus expected changes in leaching may be small and difficult to detect because of the enormous natural variability of the involved processes. In order to establish and justify nutrient reduction measures in agriculture and to monitor their effects, there is an urgent need among resource management agencies to quantify these effects despite high variations of nutrient loads.

[5] In most lakes and reservoirs, phosphorus (P) is the limiting nutrient [*Schindler*, 1978; *Fisher et al.*, 1995]. Terrogeneous dissolved P compounds are usually directly bio-available, whereas most of the particulate P settles near the inlets without participating in the lake's internal

0043-1397/05/2004WR003539$09.00

P cycling [*Gächter and Wehrli*, 1998]. Therefore this paper focuses on soluble reactive phosphorus (SRP) loads, which constitute the largest fraction of bio-available P in the catchments studied [*Gächter et al.*, 2004]. Nevertheless, the sampling strategies and estimation procedure discussed below may be applied to other P compounds, as well as other dissolved substances and particulate matter, provided that the strategies are critically evaluated and adjusted accordingly. Especially for particulate P, the individual case has to be investigated whether concentrations of particulate P exhibit a well-defined dependence on discharge, or whether their maxima are well ahead of the discharge peak [*Pacini and Gächter*, 1999].

[6] The main goal of this paper is to provide a method for isolating small effects, such as those caused by nutrient reduction measures in agriculture, from naturally occurring variations in nutrient load. In order to achieve this goal, data needs to be collected at a temporal resolution and over an interval which allows robust estimates, and a technique has to be used which results in reliable loads. In an analysis of a long-term data set from a high-altitude stream, *Robinson et al.* [2004] showed a statistically significant decline in nitrate loads. However, detecting similar trends for P loads is more difficult because they exhibit a higher discharge dependency than most other constituents, including nitrate [*Vanni et al.*, 2001; *Gächter et al.*, 2004].

[7] For optimal data collection, one has to find a compromise between high-frequency sampling required by the dynamics of the system and the limitations set by monitoring costs. Various authors have developed sampling strategies, especially stressing high temporal variations [*Robertson and Roerisch*, 1999] and the importance of sampling at times of high discharge [*Preston et al.*, 1992; *Correll et al.*, 1999; *Pacini and Gächter*, 1999]. Among the estimation techniques commonly used, the rating curve method is most appropriate for the data presented here, because the relation between discharge and P concentrations typically exhibits a strong nonlinearity. In this method, concentration is expressed as a function of discharge, i.e., $C(Q)$, by fitting an empirical curve to concentration measurements at different discharges. Besides log-log rating curves, which are commonly used, additional rating curves, which account for both point and diffuse sources [*Davis and Zobrist*, 1978], are used.

[8] The usefulness of different sampling strategies was evaluated by selecting subsamples of quasi-continuous data for estimating loads using Monte Carlo simulations [e.g., *Preston et al.*, 1989; *Coats et al.*, 2002; *Kauppila and Koskiaho*, 2003]. Similarly, in this paper, continuous data sets from three Swiss streams are evaluated. The particularity of these data sets is their high temporal resolution of up to 40 samples per day. These data sets allow (1) using Monte Carlo simulations to determine an optimized sampling strategy and load estimation technique, (2) determining differences in load estimates from spot samples versus continuous sampling over various time intervals, and (3) showing differences between catchments of different sizes and land use.

[9] In addition to applying well-established load estimation techniques to the data from the three streams, a novel approach is introduced to quantify the effects of nutrient reduction measures. The first objective of this approach is to detect small changes in long-term nutrient leaching. Therefore rating curves are calculated for different years and then applied to the discharge time series of 1 year each. This procedure permits detecting trends despite the high interannual variability of discharge. The second objective is to formulate general recommendations for designing an optimized sampling strategy that reveals even weak trends in nutrient load (i.e., typically a few percent per year). In order to fulfill this objective, the dependence of errors on monitoring duration, number of samples taken, and magnitude of trends is analyzed.

2. Methods

2.1. Studied Streams and Available Data

[10] For the present study, data from three streams, discharging into two Swiss Plateau lakes, were analyzed (Table 1). Their catchments, covering areas of 3, 7 and 42 km^2, are characterized by intensive agricultural activity (livestock and cropland), and by average precipitation of around 1200 mm yr^{-1}. The "Aabach" catchment is located in a more densely populated area ($\sim$500 inhabitants km^{-2}) and the fraction of agricultural area is smaller than in the other two catchments (Table 1).

[11] In these streams, SRP concentrations were measured along with discharge at sampling frequencies of up to 40 measurements per day (Table 1). The sampling frequencies were chosen based on the hydraulic characteristics of the catchments. For Kleine Aa and Lippenrütibach, the frequencies were chosen to resolve even short-term changes in discharge and concentrations after a rain event. For Aabach, routine measurements were done once per day only, but this stream represents the largest of the three watersheds and exhibits the longest residence time. In the streams "Kleine Aa" and "Lippenrütibach," samples were taken automatically and SRP concentrations were determined photometrically using a flow injection analyzer [*Gächter et al.*, 1996]. In Aabach, flow proportional composite samples were collected over the course of 1 day each. All of these 1 day samples were refrigerated and analyzed once per week. As sample storage may have impaired data accuracy, these data are not used for all analyses in this paper. Discharge of all three streams was monitored at calibrated limnigraph stations and was digitized at the same temporal resolution as the concentration measurements. Typical time series of discharge and concentrations of the three streams are shown in Figure 1. All data were collected in the course of catchment monitoring programs [*Gächter et al.*, 1996; P. Herzog, unpublished data, 2003; P. Niederhauser, unpublished data, 2003].

2.2. Load Estimation Using Rating Curves

[12] Over a time interval (t_0 to t_1), the load L of a constituent is given by

$$L = \int_{t_o}^{t_1} Q(t) \cdot C(t) dt, \tag{1}$$

where $Q(t)$ is discharge, and $C(t)$ is the concentration as a function of time t. Discharge is commonly measured quasi-

Table 1. Characteristics of the Streams Studied

Stream	Aabach	Kleine Aa	Lippenrütibach
Tributary to	Greifensee	Sempachersee	Sempachersee
Catchment area, km^2	41.6	6.9	3.3
Land use, %			
Forest	15	15	16
Agricultural	66	79	76
Other	19	6	8
Average discharge, $m^3\ s^{-1}$	1.0	0.13	0.04
Average response time (from discharge increase to peak)	2 days	9 hours	3 hours
Average concentration of SRP, mg m^{-3}	20	71	146
Average annual load of SRP	1.7 t yr^{-1}	0.87 t yr^{-1}	0.28 t yr^{-1}
Period of data collection	Jan. 1988–Dec. 2002	March 1993–Feb. 1994 (long-term monitoring: Jan. 1986–Dec. 2001)	Jan. 1998–Dec. 2003
Sampling interval	1 day (composite samples)	35 min (spot samples)	60 min (spot samples)
Data used for simulations	yes (except sampling duration)	yes	yes
Data used for trend detection	yes	yes (long-term monitoring data)	yes
Data used for management case study	no	yes	no

continuously and is available at high temporal resolution. C, however, is usually measured at low frequencies, and therefore some extrapolation, based on the available data, is necessary in most applications.

[13] In this paper, C denotes SRP concentration, which was parameterized as a function of discharge by least squares fitting of rating curves $C(Q)$ [*Cohn et al.*, 1989] to measured values of Q and C. A common parameterization is the log-log rating curve [e.g., *Walling*, 1977]

$$\ln C = \ln c_1 + c_2 \cdot \ln Q, \tag{2}$$

which is equivalent to

$$C = c_1 \cdot Q^{c_2}. \tag{3}$$

Often, the coefficients c_1 and c_2 are determined by linear regression from equation (2). In this case, a correction factor is necessary to account for the bias introduced by retransformation of the log-log relation [*Ferguson*, 1986]. For this paper, the coefficients c_1 and c_2 were instead determined using equation (3), by minimizing the square of the deviations using the Nelder-Mead simplex (direct search) method [*Nelder and Mead*, 1965].

[14] A log-log relation between Q and C is useful in many cases, but only empirically justified [*Cohn*, 1995]. In the streams studied, C was found to be decreasing with increasing Q in the lower flow range, and increasing at larger Q (Figure 2). Therefore an empirical function was established that represents an inverse relation between Q and C for low Q, and an increase for high Q. Such a relation was introduced by *Davis and Zobrist* [1978] for streams influenced by both point sources (e.g., sewage treatment plants) and diffuse sources (e.g., leaching from soils). Here we described this relation by

$$C = \frac{c_1}{Q} + c_2 \cdot \ln(Q + c_3) \quad (c_1 \geq 0,\ c_2 > 0). \tag{4}$$

The first term represents dilution (C decreases with increasing Q; point sources), whereas the second term delineates leaching (C increases with increasing Q; diffuse sources). Again, the parameters c_1, c_2 and c_3 were determined using least squares fitting. In this paper, both the log-log relation (equation (3)) and the inverse-log relation (equation (4)) were used as rating curves (Q given in $m^3\ s^{-1}$). Depending on the data, other parameterizations $C(Q)$ may be more appropriate (i.e., yield smaller deviations from the measured values). In Figure 2, two additional rating curves are shown:

$$C = \frac{c_1}{Q} + c_2 \cdot Q^{c_3} \tag{5a}$$

$$C = \frac{c_1}{Q} + c_2 \cdot \left(1 - e^{-Q \cdot a_3}\right). \tag{5b}$$

[15] Once the parameterization $C(Q)$ was established, instantaneous load was calculated for each time interval

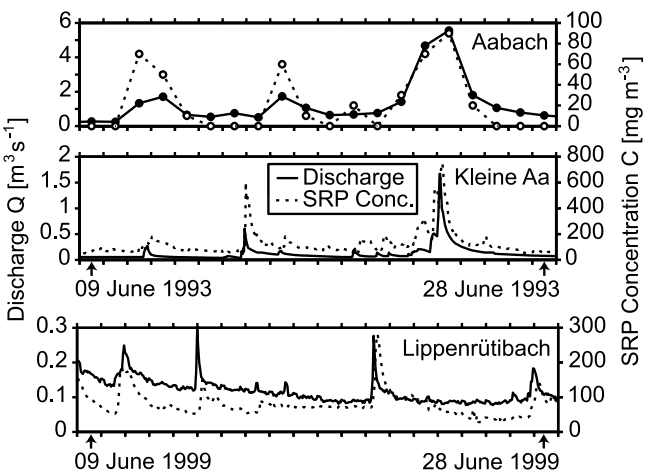

Figure 1. Time series of discharge and soluble reactive phosphorus (SRP) concentrations over a 20 day period in June 1993 (Aabach and Kleine Aa) and June 1999 (Lippenrütibach). For Aabach, daily averages are available only.

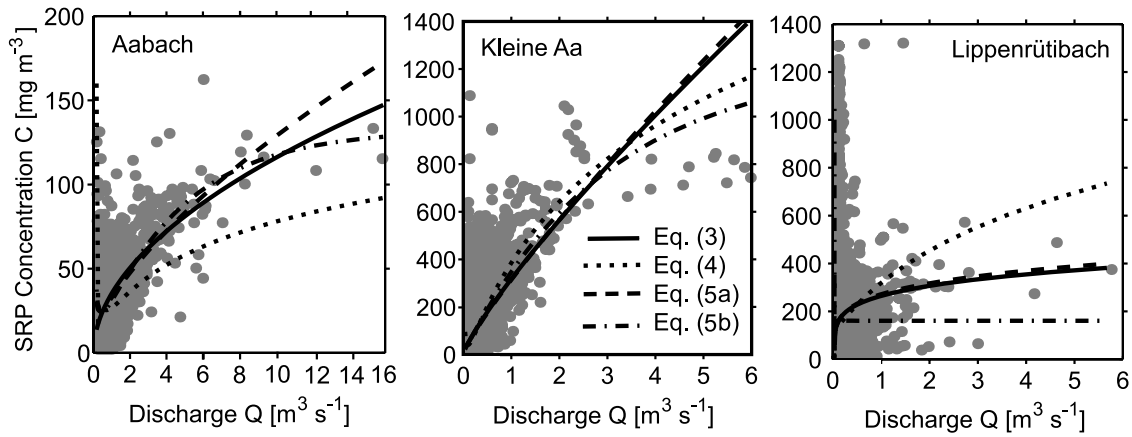

Figure 2. Measured concentration C of soluble reactive phosphorus versus discharge Q (circles). The curves result from fitting the data to four rating curves (3, 4, 5a, 5b). For equations, see section 2.2. Correlation coefficients of the log-log rating curve (equations (2) and (3)) are $R^2 = 0.50$ for Aabach, $R^2 = 0.50$ for Kleine Aa, and $R^2 = 0.33$ for Lippenrütibach.

Δt_i when Q was measured and then summed. In the following, this load will be called "estimated annual load" $\hat{L}$:

$$\hat{L} = \sum_i Q_i \cdot C(Q_i) \cdot \Delta t_i. \tag{6}$$

Q_i denotes the ith discharge measurement (average during the time interval Δt_i). For the streams discussed here, Q_i and C_i values were available simultaneously. Therefore the estimated annual load $\hat{L}$, representing an extrapolation, was compared to the reference annual load L obtained from integrating the instantaneous loads at each sampling interval Δt_i, thereby representing an approximation of the load definition in equation (1):

$$L = \sum_i Q_i \cdot C_i \cdot \Delta t_i. \tag{7}$$

2.3. Selection of Database for Load Estimation

[16] In order to design an optimized sampling strategy, the sensitivity of estimated loads to the number of samples, their distribution between times of low and high discharge, and the duration of sample collection was assessed. This evaluation was done using Monte Carlo simulations [*Preston et al.*, 1989; *Coats et al.*, 2002]. From the original Q and C records, a limited number of data pairs were chosen randomly, and $C(Q)$ parameterizations were calculated by fitting equations (3) and (4), respectively, to the selected data pairs. On the basis of these $C(Q)$ rating curves, $\hat{L}$ was estimated according to equation (6). This procedure was repeated 1000 times and allowed calculating the average estimated load $\langle \hat{L} \rangle$, its variance $\sigma^2(\hat{L})$, and the difference $(\langle \hat{L} \rangle - L)$ between $\langle \hat{L} \rangle$ and the reference load L calculated using all original data (equation (7)). Finally, mean square error (MSE), representing a superposition of the variance $\sigma^2(\hat{L})$ and of the squared difference $(\langle \hat{L} \rangle - L)^2$, was calculated according to the definition

$$MSE(\hat{L}) = \sigma^2(\hat{L}) + \left(\langle \hat{L} \rangle - L\right)^2. \tag{8}$$

In the following, we use the normalized root mean square error (NRMSE), i.e., NRMSE = $MSE^{1/2}(\hat{L})/L$.

[17] In order to select an optimized sampling strategy, simulation series were carried out by varying (1) number of data pairs, (2) number of data pairs at high Q, and (3) sampling duration, as follows.

[18] 1. In the first series, the number of data pairs used for fitting the rating curves (equations (3) and (4)) was varied between 20 and 300, to study the effect of sampling frequency on the estimated annual loads $\langle \hat{L} \rangle$ and on NRMSE.

[19] 2. In the second series, we evaluated the sensitivity of $\hat{L}$ to sampling at different discharges, especially during high-discharge events. It is evident that the distribution of the C, Q data pairs over the discharge range has a strong influence on the rating curve. Therefore the number of data points measured during high Q was varied: A total of 50 samples were taken, and a varying number n of these samples were selected from a pool of high-discharge values. This pool of high-discharge values consisted of the $10 \cdot n$ highest Q values. Such a pool of values was used in order to ensure sufficient variation in the model runs: If, for each run, the same n highest Q values had been taken, this would have resulted in artificially low variation between the runs.

[20] 3. Finally, Q and C values used for the rating curve fitting procedure were averaged over different time intervals of up to 2 days. This averaging process simulated the continuous time proportional collection of samples over the corresponding interval (composite sampling), rather than the sample being taken at a certain point in time (spot sampling). Composite sampling has the advantage of leveling out short-time fluctuations. On the other hand, if they are collected over too long intervals, they introduce bias by leveling the concentration dynamics. Composite samples were assumed to be filled at a constant rate independent of discharge. Discharge-proportional sampling (i.e., filling the sampling volume at a Q-proportional rate) would be more appropriate if loads were calculated by direct integration (equation (7)), rather than using $C(Q)$ rating curves.

2.4. Trend Estimation

[21] Variations of nutrient loads are due to both temporal variability of the meteorological boundary conditions (such as rainfall intensity and frequency) and anthropogenic

activities. From a managerial point of view, it is important to realize that changes in land use and leaching of P from soils affect the $C(Q)$ parameterization, and thus $\hat{L}$. However, such trends in P leaching are usually masked (especially if they are relatively small) by the high natural variations of Q, and therefore significant trends in $\hat{L}$ are rarely obvious.

[22] In order to remove this hydrological variability from the data, we employed the following procedure: We modeled annual P loads for a hypothetical discharge pattern, which is identical each year. There is obviously an ambiguity in the choice of such a hypothetical discharge pattern. In a first step, we described this hypothetical condition by using the discharge time series Q_i of the year of median discharge. For a later calculation (see Figure 7), we also used the time series of all other years for comparison. The index i denotes days 1 to 365 if the time series Q_i contains daily values. This discharge time series Q_i was then applied to equation (6) to calculate the hypothetical annual load $\hat{L}_j$ for year j by using the year-specific rating curves $C_j(Q)$:

$$\hat{L}_j = \sum_i Q_i \cdot C_j(Q_i) \cdot \Delta t_i. \tag{9}$$

This procedure was repeated for each year j to create hypothetical annual loads $\hat{L}_j$. Because in equation (9), rating curves $C_j(Q)$ change from year to year, whereas the discharge time series Q_i remains the same, trends in the resulting loads $\hat{L}_j$ represent trends in the rating curves $C_j(Q)$, rather than changes in discharge. Therefore the time series $\hat{L}_j$ is a way to visualize changes in nutrient leaching, and is less dependent on the hydrological situation, which typically exhibits high fluctuations from year to year.

[23] For Kleine Aa, high-resolution measurements were made over the period of 1 year only, and the time of high-resolution sampling for Lippenrütibach was 5 years only. However, additional long-term monitoring data was available from these streams: Starting in 1986, measurements of SRP were made every 22 days, with 15 to 20 additional measurements per year at high Q [*Gächter and Wehrli*, 1998]. This sampling protocol resulted in ~35 data points per year. From these long-term data sets, estimated annual loads $\hat{L}$ and hypothetical annual loads $\hat{L}_j$ were calculated as well, according to equations (6) and (9), respectively.

2.5. Error Estimation

[24] The parameter MSE (equation (8)) is an appropriate error measure for Monte Carlo simulations. Once no continuous data set is available and loads are estimated from sparse data, which was the case for the monitoring data sets used here, errors have to be calculated differently: Errors in estimated load stem from uncertainties of the measured Q and C values and the deficiency of rating curve fits. The total error of the load was calculated using systematic and statistical errors for Q (5% and 20%, respectively), statistical errors for C (5%), and the errors of the $C(Q)$ parameterization, which were calculated from the deviation of the data points from the rating curve. The total error was then obtained by linear addition of the systematic error components and quadratic addition of the statistical error components.

[25] The relation between number of samples, length of the study and trend can be expressed as follows: The error e_1 of $\hat{L}_j$ can be calculated as previously shown, with smaller errors for more samples collected. After n years, the error of the load estimation is reduced according to [*Sachs*, 1982]

$$e_2 = \sqrt{\frac{2}{n-1}} \cdot e_1. \tag{10}$$

For different combinations of number of samples and length of study, an error e_2 was calculated, using errors e_1 obtained for Kleine Aa. This error allowed determining after how many years of monitoring a trend will become apparent. Evidently, the requirement was that after n years, the trend-related change of the hypothetical load had to exceed the error e_2. This procedure allowed (for an expected trend in a given system) not only estimating the timescale of the monitoring program, but also the number of samples required.

3. Results and Discussion

[26] Figure 1 shows discharges and SRP concentrations during typical storm events. Owing to different catchment sizes, their response times and lengths of high-discharge periods vary between the three catchments. It can also be seen that concentration and discharge peaks are shifted over time in Aabach and Lippenrütibach.

3.1. Sensitivity of Load Estimates

[27] First we address the question: How many data pairs of Q and C are necessary in order to obtain a reliable estimated annual load $\hat{L}$? As detailed above, we determined $C(Q)$ rating curves for different numbers of Q,C data pairs and extrapolated to $\hat{L}$ by using equation (6). Figure 3 shows that with increasing number of Q,C data pairs used, $\hat{L}$ approaches the reference load L, calculated from the complete original data set (equation (7)), and the uncertainty, expressed by NRMSE, decreases. For Aabach and Lippenrütibach, the log-log rating curve (equation (3)) results in loads generally closer to L, but in higher uncertainty due to higher variability of the Q and C values. For Kleine Aa, the inverse-log relation (equation (4)) provides $\hat{L}$ closer to L, and with smaller uncertainties.

[28] Figure 3 gives a first insight into the effect of monitoring program design on uncertainty: The left panel shows that for Aabach about 50 Q, C data pairs have to be used for the $C(Q)$ rating curve in order to reduce the uncertainty below 10%. In Kleine Aa, using equation (3), NRMSE remains high because equation (3) cannot be fitted closely to the data points (see Figure 2 (middle)). In Lippenrütibach, more data pairs are required to reduce uncertainty because of their large scatter (Figure 2), especially in the range of low Q.

[29] In a second step, these errors can be reduced if a larger fraction of data pairs is taken during periods of high discharge. Figure 4 shows that in Aabach and Lippenrütibach, NRMSE is reduced if high-discharge Q, C data pairs constitute between 20 and 80% of the total number of data pairs. If data for the model runs are sampled randomly and independently of Q, very few data points at high discharge are included. Consequently, the $C(Q)$ parameterization differs considerably between model runs, resulting in higher NRMSE. Conversely, if mainly high-discharge data are used, the $C(Q)$ parameterization is overinfluenced by these

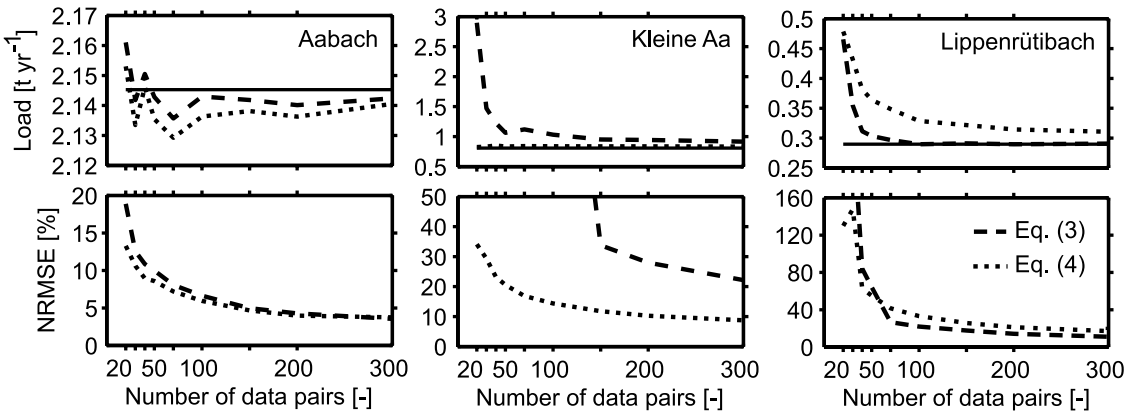

Figure 3. Estimated annual (top) SRP load $\hat{L}$ and (bottom) normalized root mean square errors (NRMSE) calculated for different numbers of Q, C data pairs using Monte Carlo simulations. The reference load L is represented by the straight line (Figure 3 (top)). Note the different scales.

data, resulting in load overestimates. For example, for Kleine Aa, $\hat{L}$ is up to 40% higher than L (Figure 4 (top middle)) if a large number of high-discharge data pairs is used. Similarly, *Robertson and Roerisch* [1999] found positive bias when the number of high-discharge samples was increased.

[30] The increase in bias for the Kleine Aa data set can be explained by the cluster of data points at discharges between 2 and 3 m^3 s^{-1} with high concentrations (Figure 2 (middle)), and can be attributed to one single high-discharge event. If a large fraction of these Q, C data pairs is used for the $C(Q)$ fitting, a steeper curve and therefore a larger $\hat{L}$ will result. This finding indicates that the $C(Q)$ rating curve is strongly affected if a cluster of data points from one single high-discharge event is included and therefore calls for sampling several high-discharge events rather than only a few. This interpretation is in line with the findings of *Richards and Holloway* [1987] that for small streams, stratified random sampling, i.e., sampling a higher proportion of high-discharge data points, is most appropriate.

[31] The effect of averaging discharge and concentration values over different time intervals (so-called "composite sampling") is shown in Figure 5. As Aabach was only sampled once per day (Table 1), these data were not used for this calculation. In Figure 5, both bias and variance are shown instead of NRMSE. This error measure helps discerning the effect of reducing scatter from the effect of adding bias. For Lippenrütibach, standard deviation $\sigma(\hat{L})$ decreases with the length of the composite sampling interval, because data scatter is reduced. On the other hand, bias increases and $\hat{L}$ are overestimated for too long intervals. This effect is critical in small catchments, where a high-discharge event can evolve within hours and is not resolved by using long sampling intervals. For the larger catchment of Kleine Aa, both standard deviation and bias decrease with the length of the sampling interval. This can also be attributed to the fact that $C(Q)$ relations in Kleine Aa are close to linear (Figure 2). If discharge and concentrations are averaged over a longer period, this hardly affects the shape of the rating curve. On the other hand, in Lippenrütibach, averaging Q and C over a longer sampling interval leads to a change in the rating curve and consequently to higher bias. Therefore in addition to the response time of the catchment (Table 1), the shape of the

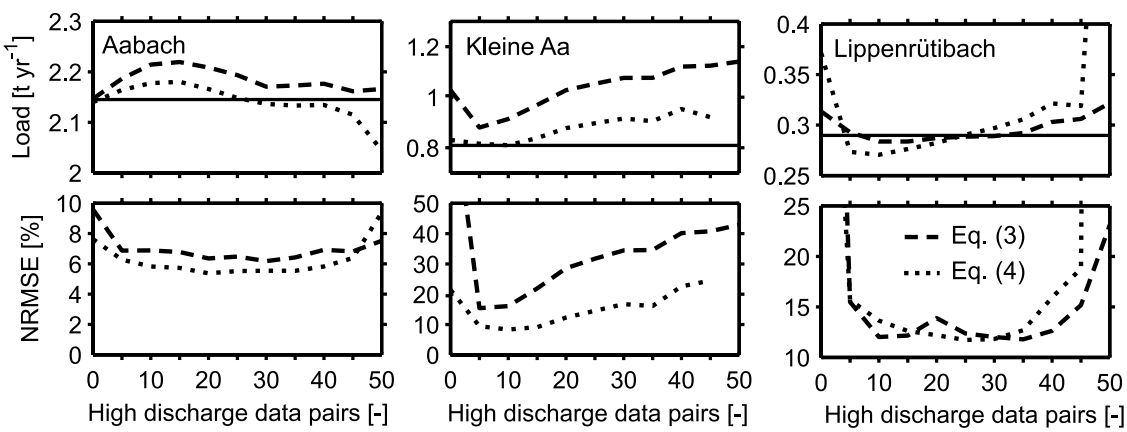

Figure 4. Estimated annual (top) SRP load $\hat{L}$ and (bottom) NRMSE using a constant number of 50 Q, C data pairs in total, with varying numbers of high-discharge Q, C data pairs. For both underrepresentation and overrepresentation of the high-discharge data pairs the uncertainty of the load estimate increases. The reference load is represented by the straight line (Figure 4 (top)). Note the different scales.

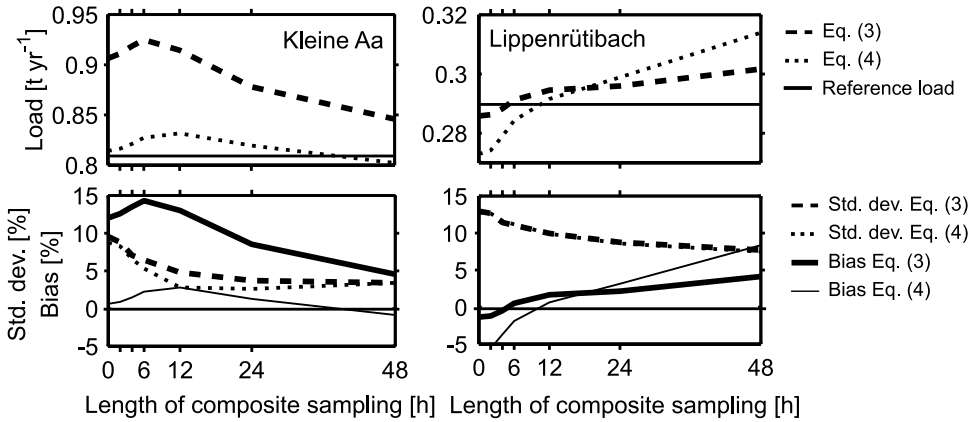

Figure 5. Estimated (top) SRP load $\hat{L}$ compared to the (bottom) reference load L, standard deviation $\sigma(L)$, and bias using 50 Q, C data pairs averaged over different time intervals. For definitions, see text.

rating curve also affects the optimal length of composite samples.

[32] The practical restriction to the composite sampling interval is given by the SRP analysis; samples may be exposed to ambient temperature for no longer than 2–3 days. Therefore collecting water samples continuously over an interval of 1 day is often appropriate, except for catchments of a few km^2 only, where shorter sampling intervals are optimal.

3.2. Trend Detection

[33] For both Aabach and Lippenrütibach, estimated annual loads $\hat{L}$ fluctuate due to varying discharge, as shown in Figure 6 (top). By contrast, if loads are calculated according to equation (9) by applying annual $C(Q)$ rating curves to the Q data of the year of maximum, median, or minimum annual discharge, the estimated loads fluctuate less (Figure 6 (bottom)). This finding demonstrates that annual fluctuations in $\hat{L}$ originate mainly from the variations in Q, whereas the effect of changing rating curves on $\hat{L}$ is much smaller. For Aabach, an upward trend is apparent for the 1990s; whereas for Lippenrütibach, no clear trend is visible for the period of the measurements.

[34] Trends in Lippenrütibach and Kleine Aa can only be seen if additional long-term monitoring data are analyzed. As described above, equation (9) was applied to monitoring data over the past 16 years from these two streams, as well as a third, neighboring stream ("Grosse Aa"), which has a larger catchment area (15.7 km^2). Figure 7 shows the resulting loads $\hat{L}$. For this figure, loads $\hat{L}$ were calculated using not only data from the year of minimum, median, and maximum discharge, but from each of the 16 discharge time series Q_i. Interestingly, all 16 curves show a similar general trend. Although total annual discharge differs considerably from year to year, applying discharge data from each of these years to changing rating curves leads to a similar trend in all cases.

[35] Figure 7 shows that although loads $\hat{L}$ (upper panel) fluctuate considerably in all three streams, loads $\hat{L}_j$ fluctuate less and for Grosse Aa and Kleine Aa show a distinct downward trend (-3% yr^{-1} on average) (Figure 7 (bottom)). For Lippenrütibach, where the high-resolution data of the past few years did not exhibit a trend, a downward trend can be seen at the end of the 1980s.

[36] The decrease in load in the three streams presented in Figure 7 may be attributed to a combination of measures such as restoration of sewer facilities, alternative fertilization schemes, or buffer strips between agricultural areas and streams. In the Aabach catchment (Figure 6), which is located in a different region, no comparable measures have

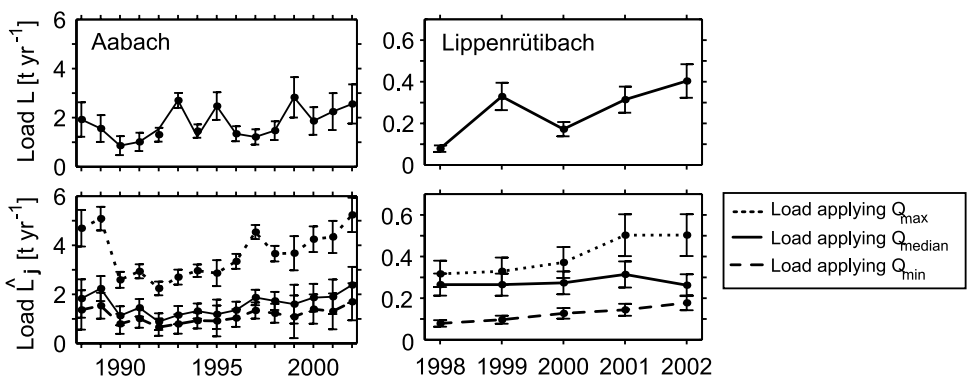

Figure 6. Annual reference SRP loads calculated using (top) original Q records and (bottom) by applying the $C(Q)$ rating curves (which change as a function of time) to three different sets of yearlong Q records. Q_{max}, Q_{median}, and Q_{min} indicate the discharge records with the maximum, median, and minimum annual discharge. Error bars denote two standard deviations.

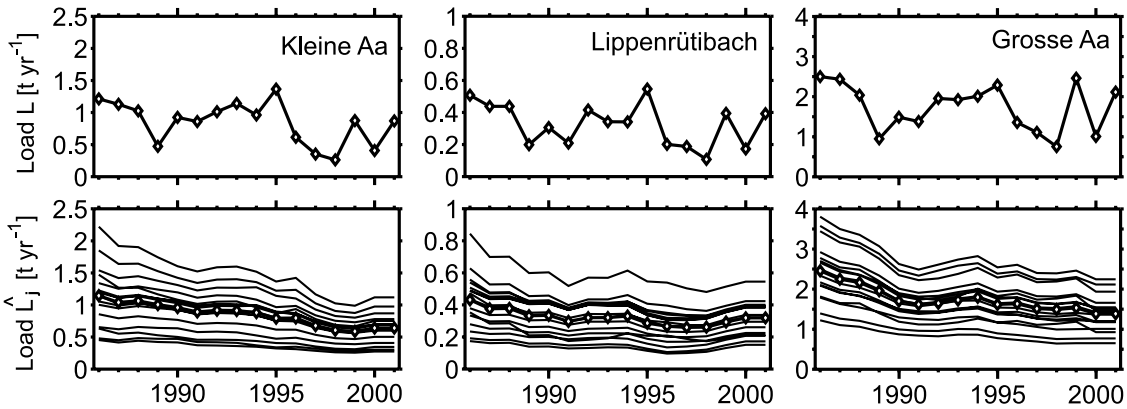

Figure 7. Annual reference loads of SRP in Kleine Aa, Lippenrütibach, and Grosse Aa, (top) calculated using all original C, Q data pairs and (bottom) loads $\hat{L}$ calculated by applying the $C(Q)$ rating curves to the discharge data sets of each individual year. Diamonds mark $\hat{L}$ obtained using the discharge data set of the year of median discharge.

been implemented so far because agricultural use is less intensive, and average nutrient concentrations are smaller (Table 1).

[37] In general, it obviously takes less time for a trend to become visible the larger the trend is. For small trends, either more samples have to be collected to improve representativeness, or the investigation has to extend over a longer period. Figure 8 shows the interrelation between number of samples, length of the study and trend for Kleine Aa (calculated using equation (10)). For instance, to show a trend in load of 3% yr^{-1} (see horizontal arrow in Figure 8), ~150 annual samples have to be collected over a period of 3 years. Alternatively, ~30 annual samples have to be taken over a period of 5 years (see vertical arrow in Figure 8). Therefore the longer the monitoring program, the fewer total resources are needed (in the example, a total of ~150 samples for the 5 year period versus a total of ~450 samples for the 3 year period). From a managerial viewpoint it is important to notice that the 5 year option reduces the annual monitoring costs by a factor of five (30 instead of 150 annual samples).

[38] Seen another way, if certain resources (e.g., 450 samples total) are available, the uncertainty can be reduced substantially if they are distributed over a longer time period. In the example, if 450 samples are spread over 5 years (i.e., 90 samples per year), trends below ~2% may still be detected. If these 450 samples are spread over a period of 3 years only, trends may be detected down to ~3% only. A graph such as Figure 8, based on high-resolution sampling done during 1 year, helps in determining an appropriate sampling strategy taking into account restrictions to sampling costs.

3.3. Discussion of the General Approach

[39] The load estimates carried out here are based on the following simplifications: (1) Seasonal variations in the $C(Q)$ parameterization were not taken into account. Depending on the catchment characteristics and land use practice, seasonal rating curves may be used [*Clement*, 2001]. However, the Q and C time series of the streams considered here suggested that there is no pronounced difference between seasons. (2) Data from Kleine Aa showed variations in concentration over the course of a day in springtime (not shown). Such variations are not taken into account explicitly, but are averaged out if composite sampling extends over the course of 24 hours. (3) *Gächter et al.* [2004] showed that SRP concentrations exhibit a hysteresis; their increase and decrease lag behind increase and decrease of Q (see also Figure 1). For the catchments studied here, this effect can be reduced by taking composite samples over the course of up to 1 day. For larger catchments, the hysteresis may result in a larger scatter when plotting Q versus C, but adds no bias, provided that samples are distributed evenly over the increasing and decreasing period of high discharge.

[40] As extreme floods are rare, data pairs associated with highest discharges often originate from one single summer high-discharge event. For example, in Figure 1 (middle) all

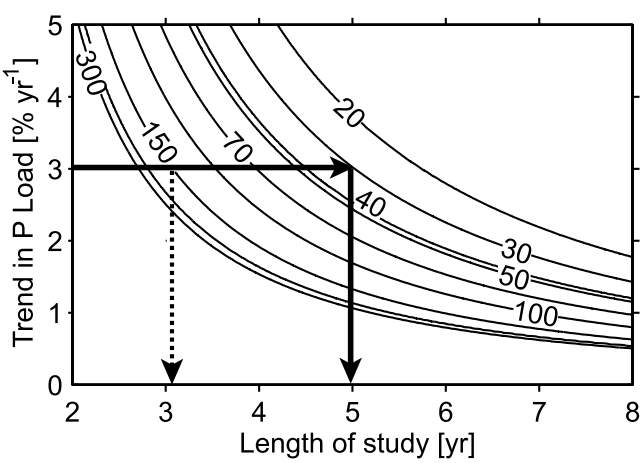

Figure 8. Number of Q, C data pairs per rating curve required for detecting a significant effect as a function of monitoring program duration and of expected trend in SRP load for the case study of Kleine Aa. The horizontal arrow indicates the observed trend (3% yr^{-1}) (see Figure 7). If 30 Q, C data pairs per year are collected for the estimation of the $C(Q)$ rating curve, then it takes 5 years (vertical arrow) until a significant trend in the annual load can be detected.

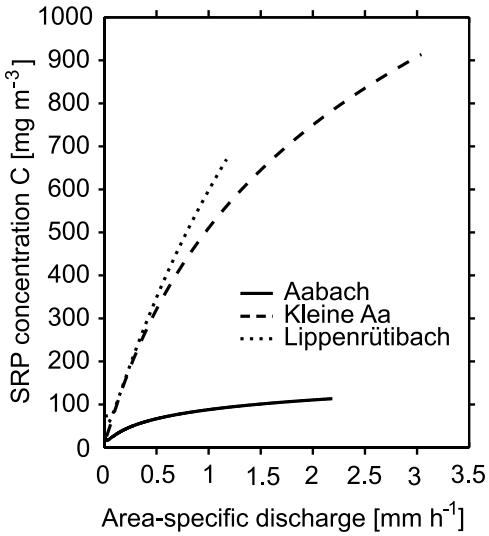

Figure 9. Measured concentration C of soluble reactive phosphorus versus area-specific discharge Q. The rating curves were calculated using the inverse-log relation (equation (4)).

data with $Q > 3$ m^3 s^{-1} are attributed to one summer flood. Because these events contribute a significant part of the annual load, it is critical to measure them. Therefore sampling protocols should provide for additional sampling of such events. A threshold for sampling, e.g., a discharge which statistically occurs four times a year, may be defined in advance. As the number of floods varies from year to year, it is important to provide flexible resources for years of multiple floods. *Richards and Holloway* [1987] showed that "fixed budget" strategies perform considerably worse than "flexible budget" strategies.

[41] In an analysis of data from eight small streams, *Robertson* [2003] elaborated how the duration of a study determines the choice of the sampling strategy. In this study, the importance of high-flow sampling was also highlighted. Similar to the present study, *Robertson* [2003] found that regressions based on about 30 points or less are generally imprecise. Concerning the problem of positive bias due to measurements at high discharge, *Robertson and Roerisch* [1999] showed that "storm chasing," i.e., measuring concentrations a short time after they reached their peak, is the most effective strategy. Alternatively, as shown in the present study, bias can be reduced by composite sampling, provided that the size of the catchment and the properties of the rating curve are taken into account (Figure 5).

[42] As far as comparison between catchments is concerned, it is instructive to use area-specific discharge (e.g., m^3 (s $\cdot$ km^2)$^{-1}$ or mm hr^{-1}) rather than discharge (m^3 s^{-1}): Figure 9 shows that Kleine Aa and Lippenrütibach exhibit much steeper rating curves, i.e., higher leaching for a given area-specific discharge. This can be attributed to the larger fraction of agricultural area and more intense fertilization application in these catchments.

4. Conclusions

[43] From the presented data analysis and the Monte Carlo simulations for optimizing sampling and load estimation strategies we conclude the following.

[44] 1. In all three streams, SRP load errors increase strongly below a threshold of around 30 to 50 Q,C data pairs, whereas for higher numbers of data pairs, errors decrease gradually only.

[45] 2. Of the two rating curves (equations (3) and (4)), both yield similar results. Differences in certain model runs suggest that for a particular stream, various rating curves should be tested, and the curve which exhibits best correspondence should be used.

[46] 3. To minimize errors in estimated annual SRP loads, routine monitoring has to be supplemented by measurements at high discharge, and composite sampling is preferable to spot sampling. For the determination of SRP loads in small catchments (<50 km^2), the sampling frequencies and composite sampling durations obtained here may serve as a guideline. For other catchment sizes or other constituents, measurements at high frequencies should be carried out, including major high-discharge events, and then long-term sampling strategies may be designed based on simulations with subsets of the collected Q, C data.

[47] 4. For optimizing sampling strategies, MSE proved to be a useful error measure when using simulations. If the scatter of data points is of main interest, variance may be used as an error measure instead. On the other hand, if the deviation from the actual load is critical (e.g., when determining the duration of composite samples (Figure 5)), bias should be used as an additional error measure.

[48] 5. For analyzing small long-term trends in nutrient leaching, we propose selecting a sampling protocol as described above, and calculating hypothetical loads as shown in equation (9) by removing hydrological variability. As illustrated in Figure 8, the selected strategy is a compromise between resources (number of samples) and duration of the monitoring program. From experience with Kleine Aa, one can conclude that at least 30 Q, C data pairs per year are necessary to show a trend on the order of 3% over the course of 5 years. Any other combination of monitoring duration, trend and number of samples can be extracted from Figure 8. In general, the longer the duration of a monitoring program, the fewer total samples are required to detect an expected trend. In our example, annual monitoring costs are reduced by a factor of five if monitoring is extended over a 5 year period instead of a 3 year period.

[49] 6. The basic concept discussed here may also be applied to other constituents like particulate P or suspended solids, provided that a $C(Q)$ rating curve can be found for the individual data set. The shape of the rating curve indicates whether the constituent exhibits more of a dilution effect or increases during high discharge. In each of these cases, the trend analysis described here (if adjusted accordingly) can be a useful tool for visualizing changes in human impact on the environment, normally disguised by high natural fluctuations.

[50] **Acknowledgments.** Data acquisition was funded by the canton of Lucerne, by the COST action 832 "Methodologies for estimation of the agricultural contribution to eutrophication," by the Swiss Agency for the Environment, Forests and Landscape (BUWAL), and by EAWAG. We are grateful to C. Crespi and A. Mares for conducting data acquisition. We thank A. Matzinger and M. Reinhardt as well as two anonymous reviewers for helpful comments and C. Hoyle and D. McGinnis for proofreading the manuscript.

References

Bodo, B., and T. E. Unny (1983), Sampling strategies for mass-discharge estimation, *J. Environ. Eng.*, *109*, 812–828.

Clement, A. (2001), Improving uncertain nutrient load estimates for Lake Balaton, *Water Sci. Technol.*, *43*, 279–286.

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

Cohn, T. A. (1995), Recent advances in statistical methods for the estimation of sediment and nutrient transport in rivers, *Rev. Geophys.*, *33*, 1117–1123.

Cohn, T. A., L. L. DeLong, E. J. Gilroy, R. M. Hirsch, and D. K. Wells (1989), Estimating constituent loads, *Water Resour. Res.*, *25*, 937–942.

Correll, D. L., T. E. Jordan, and D. E. Weller (1999), Transport of nitrogen and phosphorus from Rhode River watersheds during storm events, *Water Resour. Res.*, *35*, 2513–2521.

Davis, J. S., and J. Zobrist (1978), The interrelationships among chemical parameters in rivers—Analysing the effect of natural and anthropogenic sources, *Prog. Water Technol.*, *10*, 65–78.

Dolan, D. M., K. A. Yui, and R. D. Geist (1981), Evaluation of river load estimation methods for total phosphorus, *J. Great Lakes Res.*, *7*, 207–214.

Ferguson, R. I. (1986), River loads underestimated by rating curves, *Water Resour. Res.*, *22*, 74–76.

Fisher, T. R., J. M. Melack, J. U. Grobbelaar, and R. W. Howarth (1995), Nutrient limitation of phytoplankton and eutrophication of inland, estuarine, and marine water, in *Phosphorus in the Global Environment*, edited by H. Tiessen, pp. 301–322, John Wiley, Hoboken, N. J.

Gächter, R., and B. Wehrli (1998), Ten years of artificial mixing and oxygenation: No effect on the internal phosphorus loading of two eutrophic lakes, *Environ. Sci. Technol.*, *32*, 3659–3665.

Gächter, R., A. Mares, C. Stamm, U. Kunze, and J. Blum (1996), Dünger düngt Sempachersee, *Agrarforschung*, *3*, 329–332.

Gächter, R., S. M. Steingruber, M. Reinhardt, and B. Wehrli (2004), Nutrient transfer from soil to surface waters: Differences between nitrate and phosphate, *Aquat. Sci.*, *66*, 117–122.

Galat, D. L. (1990), Seasonal and long-term trends in Truckee River nutrient concentrations and loadings to Pyramid Lake, Nevada: A terminal saline lake, *Water Res.*, *24*, 1031–1040.

Kauppila, P., and J. Koskiaho (2003), Evaluation of annual loads of nutrients and suspended solids in Baltic Rivers, *Nordic Hydrol.*, *34*, 203–220.

Nelder, A., and R. Mead (1965), A simplex method for function minimization, *Comput. J.*, *7*, 308–313.

Pacini, N., and R. Gächter (1999), Speciation of riverine particulate phosphorus during rain events, *Biogeochemistry*, *47*, 87–109.

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

Preston, S. D., V. J. Bierman, and S. E. Silliman (1992), Impact of flow variability on error in estimation of tributary mass loads, *J. Environ. Eng.*, *118*, 402–419.

Richards, R. P., and J. Holloway (1987), Monte Carlo studies of sampling strategies for estimating tributary loads, *Water Resour. Res.*, *23*, 1939–1948.

Robertson, D. M. (2003), Influence of different temporal sampling strategies on estimating total phosphorus and suspended sediment concentration and transport in small streams, *J. Am. Water Resour. Assoc.*, *39*, 1281–1308.

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

Robinson, R. B., M. S. Wood, J. L. Smoot, and S. E. Moore (2004), Parametric modeling of water quality and sampling strategy in a high-altitude Appalachian stream, *J. Hydrol.*, *287*, 62–73.

Sachs, L. (1982), *Applied Statistics*, 706 pp., Springer, New York.

Schindler, D. W. (1978), Factors regulating phytoplankton production and standing crop in the world's freshwaters, *Limnol. Oceanogr.*, *23*, 478–486.

Thomas, R. B., and J. Lewis (1995), An evaluation of flow-stratified sampling for estimating suspended sediment loads, *J. Hydrol.*, *170*, 27–45.

Vanni, M. J., W. H. Renwick, J. L. Headworth, J. D. Auch, and M. H. Schaus (2001), Dissolved and particulate nutrient flux from three adjacent agricultural watersheds: A five-year study, *Biogeochemistry*, *54*, 85–114.

Vieux, B. E., and F. G. Moreda (2003), Nutrient loading assessment in the Illinois River using a synthetic approach, *J. Am. Water Resour. Assoc.*, *39*, 757–769.

Walling, D. E. (1977), Assessing the accuracy of suspended sediment rating curves for a small basin, *Water Resour. Res.*, *13*, 531–538.

E. Butscher and P. Herzog, Umwelt und Energie, Kanton Luzern, CH-6002 Luzern, Switzerland. (ernst.butscher@lu.ch; peter.herzog@lu.ch)

R. Gächter, L. Moosmann, B. Müller, and A. Wüest, Limnological Research Center, Swiss Federal Institute for Environmental Science and Technology, CH-6047 Kastanienbaum, Switzerland. (rene.gaechter@eawag.ch; lorenz.moosmann@eawag.ch; beat.mueller@eawag.ch; alfred.wuest@eawag.ch)