CONSIDERATIONS IN SELECTING A WATER QUALITY SAMPLING STRATEGY

K. W. King, R. D. Harmel

ABSTRACT. *Water quality monitoring programs have expanded in an effort to quantify loadings to streams and lakes from various watershed activities and managements. At the core of monitoring programs are strategies or schemes that determine when and how samples are taken for estimating stream loadings. Quantification of the differences between these schemes has not been adequately documented. An analytic approach was used to evaluate 45 commonly used sampling strategies that included time–based (5, 10, 15, 30, 60, 120, 180, 300, and 360 min) and flow–stratified (2.5, 5.0, 7.5, 10.0, 12.5, and 15.0 mm) schemes using discrete and composite sampling procedures. A total of 300 storm hydrographs from 87 different watersheds in the U.S. were coupled with two concentration graphs (a 100% positive correlation of concentration to flow, and a 100% negative correlation to flow) to estimate average bias values for each sampling strategy. The mean bias and absolute error for time–based sampling, as determined by the standardized root mean square error (SRMSE), always increased with a greater sampling time interval. For time–based sampling, a positive correlated concentration graph generally resulted in under–prediction (positive bias) from the true load, while a negative correlated concentration always resulted in over–prediction (negative bias). For flow–stratified sampling, the direction of bias was generally reversed from the time–based case, but the SRMSE increased with a greater flow interval. Even at the lowest flow interval used in this study (2.5 mm), the median residual values were significantly different from zero (α = 0.05). Time discrete sampling schemes $\leq$15–min provided the only bias and mean residual values not significantly different from zero (α = 0.05). When an equal number of samples was obtained, the flow–stratified approach had less absolute error than did the time–based approach. The results indicate that, prior to water quality monitoring, careful consideration should be given to the sampling strategy and its overall impact on load estimates.*

Keywords. *Composite sampling, Flow–stratified, Monitoring, Stream loading, Time discrete, TMDLs.*

Field/watershed–scale water quality monitoring is usually initiated with a goal of (1) collecting data for model development or validation efforts, (2) quantifying impacts of alternative management practices, or (3) measuring loads for regulatory compliance. Model developers depend on accurate field data for model development and validation, and to help define processes that are not well understood. Federal, state, and local agencies depend on monitoring programs for regulatory compliance procedures and to evaluate the effectiveness of subsequent land use changes.

Pollutant loadings to water bodies have been a global concern for some time. These loadings account for millions of dollars of damage and require even greater amounts of cleanup expense. Even though it was not until the Clean Water Act of 1972 that the concept of quantifying loads from various watershed managements and land uses was given the name "total maximum daily load" (TMDL), many studies had already been initiated in an effort to answer some of the same questions posed by that 1972 legislation (Miller, 1963; Chapman et al., 1967; Keup, 1968; Holt, 1969; Gilbertson, et al., 1971; Taylor et al., 1971; Romkens et al., 1973; Schuman et al., 1973). Even though the 1972 Clean Water Act required establishment of TMDLs, only recently have efforts regarding the implementation and compliance of TMDLs been escalated. The basic definition of a TMDL (the maximum amount of a pollutant that a water body can receive and still meet water quality standards) requires monitoring of water resources within and at the outlet of a watershed so that concentrations and loadings may be documented (Tate et al., 1999).

The concept of water quality sampling equipment originated in the 1930s and generally focused on the ability to estimate soil loss. Early sampling equipment included: total collection devices designed to collect total runoff and sediment losses from small plots, slot–type samplers (Harrold and Krimgold, 1943; Parsons, 1954; Mutchler, 1963; Dendy, 1973), non–automated suspended sediment samplers (Federal Interagency River Basin Committee, 1948, 1952), and automated suspended sediment samplers, which include both single–stage (Federal Interagency River Basin Committee on Water Resources, 1963) and pump–type samplers (Federal Interagency River Basin Committee on Water Resources, 1962; Doty, 1970; Rausch and Haden, 1974; Claridge, 1975; Allen et al., 1976; Martin and White, 1982). With the exception of cosmetic changes, compactness, portability, and electronic interfaces, which allow for a wide range of programmable sampling schemes, present day water

Article was submitted for review in February 2002; approved for publication by the Soil & Water Division of ASAE in October 2002.

The authors are **Kevin W. King, ASAE Member Engineer,** Agricultural Engineer, USDA–ARS, Columbus, Ohio, and **R. Daren Harmel, ASAE Member Engineer,** Agricultural Engineer, USDA–ARS, Temple, Texas. **Corresponding author:** Kevin W. King, USDA–ARS, 590 Woody Hayes Drive, Columbus, OH 43210; phone: 614–292–9806; fax: 614–292–9448; e–mail: king.220@osu.edu.

quality sampling devices do not vary significantly from the original automated pump–type suspended sediment samplers.

Monitoring programs generally focus on the needs for regulatory compliance and for quantifying the effectiveness of altering land–use management and activities (Tate et al., 1999). These sampling programs generally involve a wide range of sampling frequencies and designs (Richards and Holloway, 1987) that are dependent on both temporal (storm event, seasonal base flow, and annual fluctuations) and spatial (small plots, fields, watersheds, and river basins) scales. Sampling strategies are generally stratified by either time or flow and include both discrete and composite sampling. Attempts to compare different sampling schemes using a Monte Carlo approach (Richards and Holloway, 1987; Miller et al., 2000) and measured data (Stevens and Smith, 1978; Yaksich and Verhoff, 1983; Shih et al., 1994; Thomas and Lewis, 1995; Robertson and Roerish, 1999) have been accomplished on generally large watersheds.

Richards and Holloway (1987) used a Monte Carlo approach to evaluate seven sampling strategies, which included non–stratified fixed frequency and flow–stratified sampling techniques on three watersheds ranging in size from 386 to 17000 km^2. The Richards and Holloway (1987) approach provided for a maximum of four samples per day using a fixed frequency scheme. The error associated with the fixed frequency approach was attributed to inadequate measurements of large fluxes during storm flows. Flow–stratified sampling proved to better represent the true load in all tested scenarios.

Robertson and Roerish (1999) used measured data from eight agricultural watersheds (14 to 110 km^2) in Wisconsin to evaluate ten sampling strategies which include three fixed–period sampling schemes and seven fixed–period with storm chasing (a higher frequency of samples is taken during storm flow events) sampling schemes. The measured data included bimonthly grab samples and approximately 10 to 20 storm events with 6 to 10 samples per event. Robertson and Roerish (1999) concluded that the best strategy for arriving at daily loads involves storm chasing, while the least precise methods included single–stage and peak–flow sampling techniques.

Previous studies have generally focused on larger, continuous flowing streams, and the findings suggest that more intensive forms of sampling generally result in more accurate estimates of the true load. While attempts have been made to compare one, two, or even several sampling schemes to the true load, a comprehensive analytical approach that quantifies the tradeoffs between the differing schemes has not been well documented.

Discrete sampling implies the collection of one sample per bottle, while composite sampling involves the collection of more than one sample per bottle. Time–based sampling is based on a pattern of times (e.g., every 15 min), while flow–stratified sampling is based on flow past a certain point (e.g., every 2.5 mm volumetric depth over the watershed). Discrete time–based and flow–stratified approaches are illustrated in figure 1.

Objectives

With the onset of TMDL legislation, data collection efforts will continue to intensify in many watersheds. One of the primary questions that will arise out of this effort will concern the optimal sampling strategy that adheres to the constraints of economic efficiency (number of samples that have to be analyzed) and accurate estimation of the true load. This effort will try to answer that question for storm flow sampling. Specifically, the objectives of this study are to use an analytic approach to investigate and quantify the tradeoffs and impacts on loading estimates from storm flow using several common sampling strategies that include various time–based and flow–stratified discrete and composite schemes.

Methods and Procedures

Sampling Strategies

A wide range of sampling schemes was selected to properly depict the many schemes presently used. The sampling strategies include an array of discrete time–based and flow–stratified schemes as well as procedures that represent a broad spectrum of time and flow composite sampling. The time–based sampling schemes tested involved

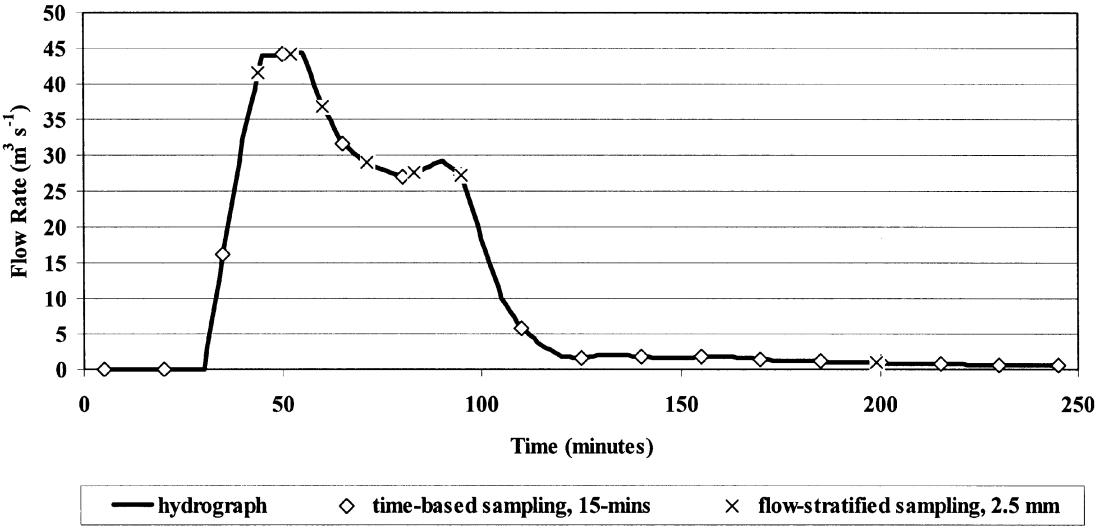


Figure 1. Example hydrograph from Walnut Gulch, Arizona (drainage area 824 ha), with discrete time–based and flow–stratified sampling strategies overlaid.

Table 1. Location of storm hydrographs.

Location	No. of Watersheds	Total No. of Events
Branch, Arkansas	1	2
Caulksville, Arkansas	1	4
Chisomville, Arkansas	2	10
Safford, Arizona	1	1
Walnut Gulch, Arizona	4	7
Ralston Creek, Iowa	1	5
Treynor, Iowa	1	1
Reynolds Creek, Idaho	1	9
Horsepen Creek, Louisiana	1	1
Creek Draw, Mississippi	1	1
Long Creek, Mississippi	1	1
Oxford, Mississippi	7	21
Short Creek, Mississippi	1	3
White Oak Creek, Mississippi	1	1
Ahoskie, North Carolina	3	6
Hastings, Nebraska	4	24
Coshocton, Ohio	6	21
Big Creek, Oklahoma	1	1
Canyon View Creek, Oklahoma	1	1
Chickasha, Oklahoma	4	6
Clear Creek, Oklahoma	1	1
Pine Creek, Oklahoma	1	1
Rock Creek, Oklahoma	1	1
Stidham Creek, Oklahoma	1	1
Stillwater, Oklahoma	1	4
Browns Creek, Tennessee	1	1
Coon Creek, Tennessee	1	1
Red River, Tennessee	1	1
Richland Creek, Tennessee	1	2
Wartrace Creek, Tennessee	1	1
Calveras, Texas	1	1
Cow Bayou, Texas	1	6
Deep Creek, Texas	2	7
Elm Fork, Texas	1	3
Escondido Creek, Texas	1	2
Green Creek, Texas	1	5
Honey Creek, Texas	2	15
Keegans Bayou, Texas	1	2
Little Elm Creek, Texas	1	8
Little Fossil Creek, Texas	1	1
Mukewater Creek, Texas	1	4
North Creek, Texas	1	6
Pin Oak Creek, Texas	1	5
Little Pond Creek, Texas	1	2
Riesel, Texas	7	70
Sims Bayou, Texas	1	3
Wilbarger Creek, Texas	1	4
Brush Creek, Virginia	1	1
Foster Creek, Virginia	1	3
Rocky Run, Virginia	1	2
Little Winns Creek, Virginia	1	2
North Danville, Vermont	1	1
Colby, Wisconsin	1	2
Fennimore, Wisconsin	1	2
La Crosse, Wisconsin	2	4

discrete sampling at 5, 10, 15, 30, 60, 120, 180, 300, and 360 min intervals. Time–based composite samples (3 and 6 samples per bottle) were also investigated using the same time intervals. For example, a 3–sample, 5–min, time–based composite sample would represent 15 min of flow with one concentration calculated as the average concentration from samples pulled at 5, 10, and 15 min. Discrete flow–stratified sampling was based on flow increments of 2.5, 5.0, 7.5, 10.0, 12.5, and 15.0 mm. A flow–stratified composite approach (3 and 6 samples per bottle) using the same flow increments was also completed. For example a 3–sample, 2.5 mm, flow–stratified composite sample would represent 7.5 mm of volumetric depth, and the concentration would be the average of the 2.5, 5.0, and 7.5 mm point concentrations. These schemes allowed the evaluation of 45 different sampling strategies.

Table 2. Watershed area, runoff amount and duration, and peak flow distribution statistics for 300 hydrographs used in the study.

Statistic	Drainage Area (ha)	Runoff Amount (mm)	Runoff Duration (min)	Peak Flow ($m^3 s^{-1}$)	Time to Peak (min)
Mean	1228	31.8	1205	24.4	322
25th percentile	234	13.2	560	3.9	120
Median	596	26.1	850	14.6	200
75th percentile	1368	47.0	1485	30.3	375
Standard deviation	1494	26.6	1002	31.8	368
Minimum	0.1	0.01	75	0.003	13
Maximum	6294	134.2	6180	202.5	2936

Analytic Approach

Hydrographs

Data from 300 storm events over an array of locations in the U.S. (table 1) were used as input for runoff hydrographs so that several hydrograph shapes could be evaluated. Flow between hydrograph points was assumed to be linear, and interpolation between the points was completed to form 1–min hydrographs. The 1–min hydrographs were used to approximate the true discharge volume. The shapes of the hydrographs varied from short–duration high peaks to long–duration low peaks, with several hydrographs having double peaks as well as those characterized by long–duration rising limbs (table 2).

Concentration Graphs

One of the difficulties in using an analytical approach to achieve the objectives of this study is developing a concentration graph. Many important water quality parameters, including sediment, nitrate, dissolved phosphorus, etc., correlate well with discharge but can take on an infinite number of shapes. However, regardless of the shape, the graph will be correlated with the hydrograph in the range of –1 to +1. Therefore, a hypothetical concentration graph corresponding to 100% positive correlation and 100% negative correlation of concentration to the hydrograph was assumed. The concentration graphs were scaled such that the maximum concentration corresponding to the peak flow in the 100% positive correlation case was equivalent to unity, while the value corresponding with the baseline was zero. Conversely, in the 100% negative correlation case, the concentration associated with the peak flow was zero, and the concentration associated with the baseline was unity. This approach allows for an envelope to be formed that encloses the maximum expected bias. Shih et al. (1994) used a similar approach but only evaluated a limited set of time composite samples.

Load Estimates

The hydrographs were coupled with the concentration graphs to calculate load estimates. The 1–min hydrographs along with the corresponding 1–min concentration graphs were assumed to represent the true load. The flow volume associated with each concentration was calculated from the midpoint of the preceding time/flow interval to the midpoint of the following time/flow interval assuming a linear relationship between hydrograph points. The load for each sampling point was calculated as the concentration associated with that point multiplied by the flow during the sampling interval. The summation of the incremental loads was assumed to represent the true load.

Statistical Analysis

A combination of two statistics and a statistical test was selected to evaluate and compare the sampling strategies. The two selected statistics were bias and standardized root mean square error (SRMSE), and the test was the one–sample Wilcoxon signed rank test.

For each hydrograph, the bias in load estimates was determined for both the positive and negative correlation cases when applicable. In short–duration or low–runoff events, the longer sampling duration or larger volume flow–stratified techniques may not be applicable. When the sampling schemes were not applicable, those hydrographs were omitted from the analysis. Percent bias as defined by Shih et al. (1994) is:

$$\beta_p = \frac{(L - \hat{L})}{L} 100 \qquad (1)$$

where

β_p = percent bias
L = "true load"
$\hat{L}$ = estimated load.

Bias provides information on the over–prediction (negative bias) and under–prediction (positive bias) for each hydrograph. The mean bias values for each sampling strategy were compared to zero (a true load bias) with a one–sample t–test (α = 0.05). However, bias is a non–symmetric statistic (bias can only go to 100% in the positive direction but can theoretically approach infinity in the negative direction), and interpretation of results based solely on bias when both positive and negative biases are present could be misleading.

To complement the use of bias, the SRMSE statistic was chosen. SRMSE can be defined as:

$$\text{SRMSE} = \frac{\left[\dfrac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}\right]^{0.5}}{\dfrac{\sum_{i=1}^{n} x_i}{n}} \qquad (2)$$

where

y_i = ith estimated value
x_i = ith true value
n = number of data pairs.

The SRMSE allows a term–by–term comparison and provides a non–dimensional estimate of the accuracy of the estimated load obtained by using a selected strategy compared to the true load. SRMSE is also a non–symmetric statistic but facilitates explanation when both over– and under–prediction are present.

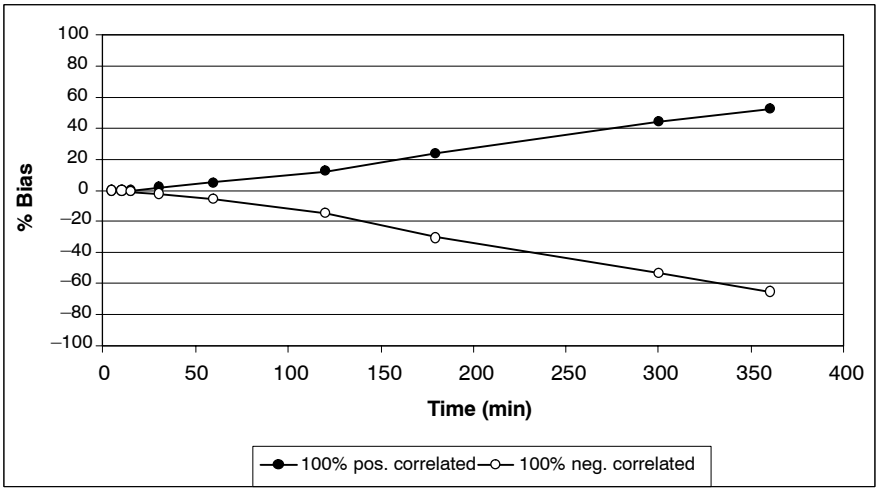
The Wilcoxon signed rank test is a non–parametric statistic of the median. Once the distribution of the residuals between the true loads and estimated loads was determined not to be normal, the Wilcoxon test was performed for each of the 45 sampling strategies using both positive and negative correlation cases to determine whether the median residual value was different from zero (α = 0.05). The Wilcoxon statistic assumes symmetry around the median of the data.

Results

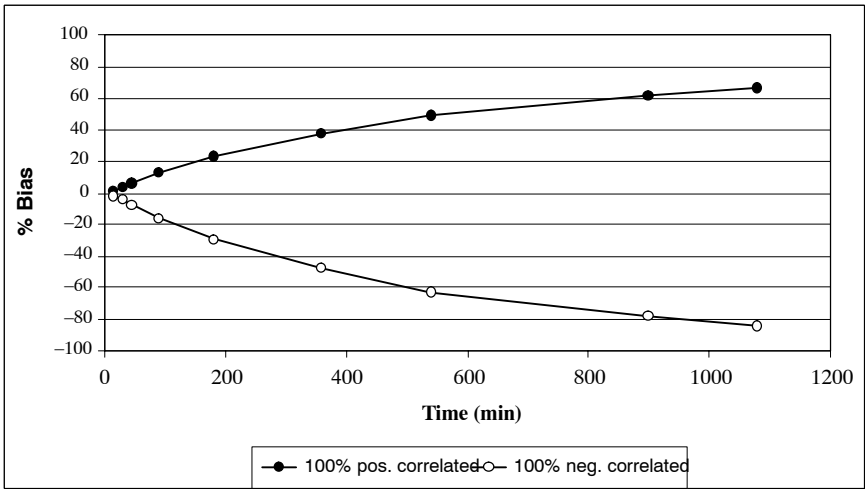
Time–Based Sampling

Assuming continuous sampling (i.e., ignoring the physical constraints of a maximum 24 bottles per sampler prior to changing bottles), the average bias associated with each sampling scheme increases with time between samples (fig. 2). The average bias also increases when moving from discrete sampling to composite sampling. A similar escalation in average bias is observed when the number of samples that are composited is increased. While compositing permits a longer runoff event to be monitored, the tradeoff is a less accurate representation of the true load. For example, if a storm event is monitored at a time discrete frequency of 15 min, then the maximum expected average bias (under–prediction, 100% positive correlation of concentration graph with hydrograph) is 0.4%, while the minimum expected average bias (over–prediction, 100% negative correlation of concentration graph with hydrograph) is –0.7% (table 3). If a composite sampling approach (3 samples per bottle) is used for that same event, then the positive bias could be expected to increase to 5.8%, while the negative bias decreases to –7.2%. A composite approach of 6 samples per bottle further increases the absolute bias. An analysis of the SRMSE for time–based sampling yields similar results (table 4). For the same 15–min time discrete sampling scheme, the SRMSE is 0.41% for the positive correlation case and 0.57% for the negative correlation case. The similarity in results is an indication that most loads were consistently under–predicted for the positive correlation and over–predicted for the negative.

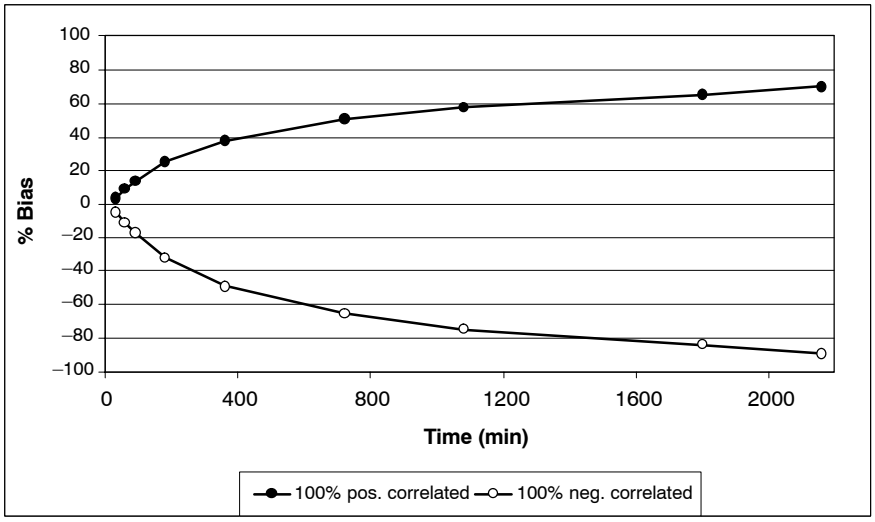
One common constraint of monitoring is economic feasibility (to use as few of samples as possible and still represent the true load). To evaluate the measured load for each sampling scheme against the true load, an assumption of normality was made, and the mean bias estimates were compared to zero (true load bias value). In the case of time discrete sampling, the bias becomes statistically different (α=0.05) from zero at time intervals greater than 15 min (table 3). Using a composite approach of 3 and 6 samples per bottle results in all bias values being significantly different from zero. The normality assumption was tested using the Anderson–Darling test and found to be false. The Wilcoxon non–parametric test was then run on the residuals of the true load and estimated loads for each sampling scheme (table 4). The median residual for both 10– and 15–min time discrete sampling schemes was not significantly different (α = 0.05)



(a)



(b)



(c)

Figure 2. Mean percent bias in the load calculation from (a) time discrete samples, (b) time composite samples (3 samples/bottle), and (c) time composite samples (6 samples/bottle).

Table 3. Statistical comparison of the mean percent bias with respect to zero for each time–based sampling strategy.

Sample Interval (min)	No. of Events	Time Discrete		Time Composite (3 samples)		Time Composite (6 samples)	
		100% Pos. Correlation	100% Neg. Correlation	100% Pos. Correlation	100% Neg. Correlation	100% Pos. Correlation	100% Neg. Correlation
5	300	0.09 N[a]	–0.25 N	1.25 Y	–1.64 Y	3.62 Y	–4.58 Y
10	300	–0.05 N	–0.12 N	3.16 Y	–4.09 Y	8.85 Y	–11.1 Y
15	300	0.40 N	–0.74 N	5.78 Y	–7.24 Y	13.6 Y	–16.9 Y
30	300	1.80 Y	–2.24 Y	12.9 Y	–15.8 Y	24.8 Y	–31.5 Y
60	300	4.76 Y	–5.37 Y	23.2 Y	–29.2 Y	37.7 Y	–48.4 Y
120	299	12.5 Y	–14.5 Y	37.3 Y	–47.3 Y	50.6 Y	–64.9 Y
180	299	23.9 Y	–30.0 Y	48.7 Y	–62.9 Y	57.7 Y	–74.6 Y
300	285	43.2 Y	–52.6 Y	60.9 Y	–77.1 Y	65.1 Y	–82.9 Y
360	271	52.3 Y	–65.3 Y	66.2 Y	–84.1 Y	69.8 Y	–88.8 Y

[a] Y = mean bias in load estimates for that sampling scheme is significantly different from zero; N = mean bias in load estimates for that sampling scheme is not significantly different from zero using a t–test ($\alpha = 0.05$).

Table 4. Statistical analysis including estimated median residuals, associated P–values, and standardized root mean square error (SRMSE) of time–based sampling strategies.

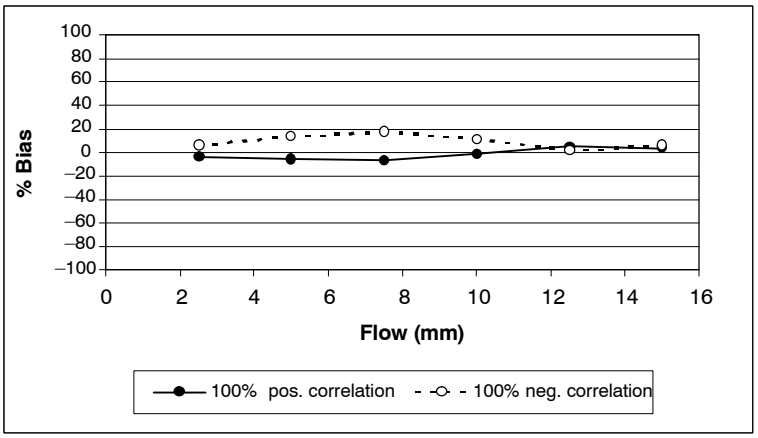
Sampling Strategy	Positive Correlation			Negative Correlation		
	Estimated Median Residual	P–value[a]	SRMSE (%)	Estimated Median Residual	P–value[a]	SRMSE (%)
Time discrete						
5 min	–0.018	0.000	0.04	0.018	0.000	0.06
10 min	0.008	0.112	0.26	–0.009	0.092	0.36
15 min	–0.029	0.057	0.41	0.031	0.044	0.57
30 min	0.241	0.000	2.43	–0.235	0.000	3.37
60 min	1.51	0.000	8.52	–1.45	0.000	11.8
120 min	2.83	0.000	15.5	–2.67	0.000	21.6
180 min	10.5	0.000	29.4	–10.4	0.000	40.4
300 min	26.9	0.000	54.7	–26.2	0.000	75.6
360 min	39.9	0.000	64.5	–39.8	0.000	88.8
Time composite (3 samples)						
5 min	0.36	0.000	0.53	–0.36	0.000	0.74
10 min	1.35	0.000	1.74	–1.35	0.000	2.41
15 min	2.70	0.000	3.43	–2.69	0.000	4.76
30 min	8.23	0.000	11.1	–8.22	0.000	15.4
60 min	19.2	0.000	22.1	–19.1	0.000	30.6
120 min	33.9	0.000	39.8	–33.9	0.000	55.1
180 min	44.9	0.000	55.6	–44.5	0.000	76.9
300 min	60.5	0.000	78.8	–59.5	0.000	109
360 min	70.6	0.000	87.6	–70.1	0.000	121
Time composite (6 samples)						
5 min	1.47	0.000	2.02	–1.47	0.000	2.8
10 min	4.74	0.000	5.34	–4.74	0.000	7.4
15 min	8.62	0.000	11.1	–8.62	0.000	15.4
30 min	21.4	0.000	22.5	–21.4	0.000	31.2
60 min	36.4	0.000	44.2	–36.2	0.000	61.4
120 min	53.2	0.000	66.9	–52.9	0.000	92.9
180 min	60.0	0.000	8.2.2	–59.6	0.000	114
300 min	71.5	0.000	97.8	–70.9	0.000	136
360 min	80.3	0.000	101	–79.8	0.000	141

[a] P–values < 0.05 indicate that the median residual value is significantly different from zero using the non–parametric Wilcoxon Signed Rank test.

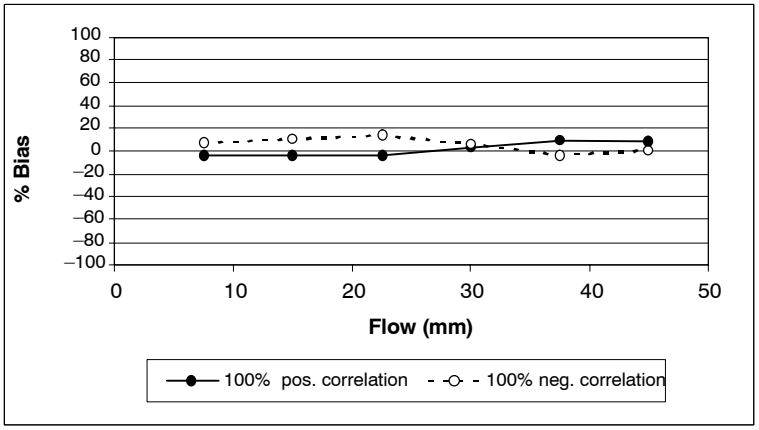
from zero for the positive correlation. In the case of negative correlation, only the 10–min time discrete strategy was not significantly different from zero. In both cases, the median residual associated with the 5–min time discrete strategy was significantly different from zero, which could only be explained by the small standard deviation (0.09) of the residuals for that strategy and the fact that the symmetry assumption was not true. Using the previous example of discrete versus composite sampling, it can be concluded that the 15–min discrete samples would represent the true load (bias not significantly different from zero), while the composite approach of 3 and 6 samples/bottle would be significantly different from the true load.

Flow–Stratified Sampling

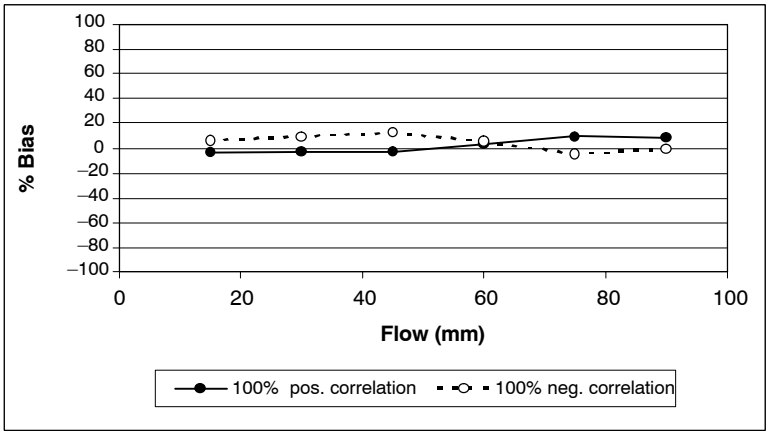
The average absolute bias associated with flow–stratified sampling generally increased with greater flow interval but then began to converge past a certain point, most likely due to a combination of a reduced number of samples, lower number of storms available for analysis, and cancellation of



(a)



(b)



(c)

Figure 3. Mean percent bias in the load calculation from (a) flow discrete samples, (b) flow composite samples (3 samples/bottle), and (c) flow composite samples (6 samples/bottle).

biases due to similar magnitude positive and negative biases (fig. 3). Assuming continuous sampling, average bias from flow–stratified discrete sampling initiated at 2.5 mm was generally negative (over–prediction) for the 100% positive correlated case and always positive (under–prediction) for the 100% negative correlated case (fig. 3a). Compositing the samples tended to reduce the level of absolute bias. For example, if a flow–stratified sampling scheme initiated at 5 mm runoff were used, the average bias from a 100% positive correlated concentration graph would be –5.4%, while the average bias associated with a 100% negative correlated concentration graph would be 12.9% (table 5). Composite sampling (3 samples per bottle) reduces the average absolute bias to –3.6% (positive correlated con–

Table 5. Statistical comparison of the mean percent bias with respect to zero for each flow–stratified sampling strategy.

Sample Interval (mm)	No. of Events	Flow Discrete		Composite (3 samples)		Composite (6 samples)	
		100% Pos. Correlation	100% Neg. Correlation	100% Pos. Correlation	100% Neg. Correlation	100% Pos. Correlation	100% Neg. Correlation
2.5	258	–3.57 Y[a]	6.63 Y	–3.97 Y	7.30 Y	–3.46 Y	6.73 Y
5.0	255	–5.35 Y	12.9 Y	–3.60 Y	10.8 Y	–2.49 Y	9.54 Y
7.5	251	–5.90 Y	16.8 Y	–2.82 Y	13.0 Y	–1.64 Y	11.7 Y
10.0	245	–1.25 N	10.9 Y	2.55 N	6.49 Y	3.52 Y	5.39 Y
12.5	236	4.44 Y	1.38 N	8.96 Y	–3.82 N	9.58 Y	–4.58 N
15.0	214	4.04 N	5.52 N	8.46 Y	0.31 N	8.84 Y	–0.26 N

[a] Y = mean bias in load estimates for that sampling scheme is significantly different from zero; N = mean bias in load estimates for that sampling scheme is not significantly different from zero using a t–test ($\alpha = 0.05$).

Table 6. Statistical analysis including estimated median residuals, associated P–values, and standardized root mean square error (SRMSE) of flow–stratified sampling strategies.

Sampling Strategy	Positive Correlation			Negative Correlation		
	Estimated Median Residual	P–value[a]	SRMSE (%)	Estimated Median Residual	P–value[a]	SRMSE (%)
Flow discrete						
2.5 mm	–2.75	0.000	3.98	4.93	0.000	9.06
5.0 mm	–7.04	0.000	10.4	10.7	0.000	22.7
7.5 mm	–10.3	0.000	13.8	17.9	0.000	31.8
10.0 mm	–9.36	0.000	20.1	16.5	0.000	43.8
12.5 mm	–6.62	0.000	27.6	12.8	0.000	51.1
15.0 mm	–6.01	0.000	29.3	15.4	0.000	56.2
Flow composite (3 samples)						
2.5 mm	–4.36	0.000	4.77	6.48	0.000	10.4
5.0 mm	–6.83	0.000	10.1	10.5	0.000	22.7
7.5 mm	–7.66	0.000	13.2	15.0	0.000	30.7
10.0 mm	–4.60	0.000	18.5	10.6	0.000	41.4
12.5 mm	–0.97	0.358	26.4	5.86	0.000	49.0
15.0 mm	–0.26	0.824	28.9	8.64	0.000	54.7
Flow composite (6 samples)						
2.5 mm	–4.52	0.000	4.77	6.64	0.000	10.5
5.0 mm	–5.68	0.000	9.26	9.25	0.000	21.7
7.5 mm	–5.82	0.000	12.4	12.7	0.000	29.8
10.0 mm	–2.89	0.000	17.6	8.56	0.000	40.4
12.5 mm	0.29	0.766	25.7	4.29	0.004	47.4
15.0 mm	0.35	0.716	28.4	7.50	0.000	53.5

[a] P–values < 0.05 indicate that the median residual value is significantly different from zero using the non–parametric Wilcoxon Signed Rank test.

centration) and 10.8% (negative correlated concentration). Understanding the convergence of bias can be facilitated by examining the SRMSE. Since the convergence is in part a result of both positive and negative biases, the SRMSE will eliminate that concern. The SRMSE associated with flow discrete sampling consistently increased for both positive and negative correlated cases (table 6). The SRMSE for the 5–mm flow discrete example used previously is considerably larger than the absolute mean bias values (both positive and negative cases) for the same strategy, which indicates a more equal distribution of biases for over– and under–prediction. The nearly equal SRMSE for both (3 and 6 samples) composite sampling schemes should be expected since the amount of discharge is consistent.

Assuming normality and comparing the average bias values to the true load bias (zero) resulted in most strategies being significantly different ($\alpha = 0.05$) from zero (table 5). Those strategies not significantly different from zero were most likely a result of the nearly equal distribution in magnitudes of positive and negative biases. A test of the normality assumption using the Anderson–Darling test revealed a non–normal distribution. As with the time–based strategies to account for the non–normality, the Wilcoxon non–parametric test was run on the median residual values for each strategy and compared to zero (table 6). With the exception of the 12.5 and 15.0 mm, 3– and 6–sample composite strategies, all median residual values were significantly different from zero ($\alpha = 0.05$). These findings suggest that on average for the watersheds used in this study, an interval smaller than 2.5 mm is needed to capture the true load.

DISCUSSION

As a result of the various storm distribution characteristics and the time and flow intervals selected, direct comparisons between time–based and flow–stratified sampling is difficult. Of the 300 events used in this study, more events had runoff durations adequate to test time–based sampling in the range of 5 to 360 min than had runoff volumes available to sample in the range of 2.5 to 15 mm volumetric flow depths for flow–stratified sampling. The number of events available

Table 7. Maximum, minimum, mean, and median number of samples taken for each time and flow discrete sampling strategy[a] for the 300 events used.

Sampling Strategy	Maximum Number of Samples	Minimum Number of Samples	Mean Number of Samples	Median Number of Samples
Time discrete				
5 min	1237	8	234	164
10 min	619	4	117	82
15 min	413	3	78	55
30 min	207	2	39	28
60 min	104	0	20	14
120 min	52	0	10	7
180 min	35	0	6	5
300 min	21	0	4	3
360 min	18	0	3	3
Flow discrete				
2.5 mm	53	0	12	10
5.0 mm	26	0	6	5
7.5 mm	17	0	3	3
10.0 mm	13	0	2	2
12.5 mm	10	0	2	2
15.0 mm	8	0	1	1

[a] Number of samples for the composite schemes can be determined by dividing the presented sample number by the number used in the composite (3 or 6 samples).

for the time–based schemes ranged from the total available 300 events at a 5–min sampling scheme to 271 for the 360–min scheme (table 3). The number of storms with adequate flow for sampling at the 2.5–mm flow–stratified scheme was 258 and decreased to 214 at the 15–mm flow interval (table 5). The number of samples taken also varied with the differing schemes (table 7). For a 59.3 mm runoff event with a duration of 1185 min, the number of time–based samples ranged from 235 for the 5–min strategy to 4 for the 360–min scheme. For the same event, the number of flow–stratified samples taken was 23 when initiated at 2.5 mm and decreased to 3 at a 15–mm flow interval. When an approximate number of samples was used to arrive at the load estimates (120–min time discrete, and 5–mm flow discrete), the flow–stratified approach showed a marked improvement in absolute bias and SRMSE, which should be expected since more samples are taken during periods of higher flows.

The average bias results for time–based sampling were consistent with expectations, namely that the positive correlated case would be under–predicted (positive bias) and the negative correlated case over–predicted (negative bias). The larger the time interval the greater the chance of missing the values around the peak flow, which in predominantly non–point source watersheds often corresponds to the time of greatest concentration (Richards and Holloway, 1987). A smaller time interval allows for more data around the peak flow but requires more sample analysis.

Unlike time–based sampling, the results of flow–stratified sampling indicate that the average bias associated with a positive correlation was generally negative (over–prediction), while the bias associated with a negative correlation was usually positive (under–prediction). These results imply that the concentration associated with the flow interval is an overestimation of the average concentration for that interval.

In both cases (time and flow), the non–symmetry associated with bias is evident. This result can be attributed to the differing magnitudes of true loads (one for the negative correlated case and one for the positive correlated case). In most cases, the true load for the negative correlated approach is much less than the true load for the positive correlated concentration approach. Even though the magnitudes of over–prediction and under–prediction are similar, the percent bias in the negative correlated case is more pronounced because of the smaller true load value. Thus, the absolute average bias associated with the negative correlated case should be expected to be greater than the positive correlated case regardless of sampling scheme, which is evidenced by examination of the SRMSE for each scheme.

With respect to flow–stratified sampling, the 7.5 mm amount appears to be the threshold interval for the storms used in this analysis where the peaks are adequately captured. Once the interval exceeds 7.5 mm, the average bias curves start to converge. The convergence, as previously noted, results from a combination of the sample interval being too large, fewer number of storms with adequate flow for analysis, and equal distribution of total magnitudes of biases. Based on the bias, SRMSE, and Wilcoxon tests, a flow–stratified interval less than 2.5 mm (the minimum flow discrete amount analyzed in this study) is needed to statistically preserve the true load.

Advantages and Disadvantages

Depending on the field/watershed to be monitored and the economic and physical constraints of the monitoring program, each strategy has advantages and disadvantages. Knowledge of the watershed runoff characteristics such as average annual runoff volume and general runoff durations will aid in selecting a sampling strategy.

Time–based discrete sampling is simple since time is easy to measure. However, for a small sampling interval (in this study shown to be approximately 15 min or less to preserve the true load), the number of samples will generally be large, limiting the sampling of storms with large runoff durations. In addition, to convert the concentrations to loads, a cross–sectional area will need to be measured and a stage discharge curve developed. In many remote locations, this cross–sectional area often changes with time. However, if concentrations are sufficient based on monitoring objectives, then the need for discharge amounts is irrelevant.

A major advantage of flow–stratified discrete sampling is more frequent sampling during high flows. The disadvantage of a flow–stratified approach is the requirement of a true control volume to measure discharge (flow interval) and the ability to continuously monitor the stage. Unlike the time–based scheme, even if concentrations are sufficient, a control volume is needed to pace the flow sampling. This requirement is often cost prohibitive or not feasible due to a remote location. In addition, during large magnitude runoff events, the number of samples allowed by the sampler prior to emptying and resetting may be exceeded. If the samples cannot be emptied and transferred, then the full event may not be sampled.

Composite sampling offers an economic advantage in that fewer samples are analyzed, permitting longer duration and larger magnitude events to be sampled. One drawback of composite sampling is the difficulty in associating the concentration with flow, especially in time–based sampling

schemes. Most importantly, as shown in this study, composite sampling with respect to time actually increases the relative and absolute bias. With respect to flow–stratified composite sampling, the absolute bias (SRMSE) does not change, which should be expected since flow is constant and the only averaging occurring is over the concentration values. In addition to the already noted negatives of composite sampling is the lack of knowledge on contaminant distribution throughout the sampling event.

Study Limitations

The results are a compilation of 300 separate runoff events on 87 watersheds with an assumption that continuous sampling (ignoring bottle number constraint) was possible. The statistical results are based on 300 events not filtered by watershed characteristics such as watershed size, time of concentration, or probable maximum peak flow. The inclusion of these characteristics was outside the scope of this study but should be expected to enhance the findings of this study. The continuation of this study will focus on determining a relationship between watershed characteristics and sampling strategies to determine a best sampling scheme for minimizing error while maximizing efficiency (small number of samples to be analyzed).

Summary and Conclusions

Water quality concerns, TMDLs, and water management planning have expanded field and watershed monitoring efforts. Water quality monitoring programs typically rely on automated samplers for collecting storm runoff samples. The sampling devices are usually programmable and allow for a wide array of sampling schemes that include discrete and composite time–based and flow–stratified approaches. Forty–five different sampling strategies were evaluated using both a 100% positive correlated concentration graph to hydrograph and a 100% negative correlated concentration graph to hydrograph to encompass the range of possible values. The resulting load estimates were compared to the true load using bias, SRMSE, and a one–sample Wilcoxon signed rank test. Mean bias values and median residual values were statistically compared to zero for each sampling strategy.

Results from the analysis of time–based strategies indicate that smaller time intervals provide smaller mean bias and SRMSEs, and preserve the true load representation. However, to sample entire storms, increased time intervals or composite sampling schemes, which have been shown to be significantly different from the true load, are typically used. Mean bias and median residual estimates associated with the larger time discrete (>15–min) and all time composite strategies were different from zero, an indication that those schemes do not represent the true load.

Calculated mean biases from the flow–stratified approach were coupled with SRMSE to more accurately understand the load estimates for each strategy. The absolute error derived from SRMSE increased consistently with a larger flow interval. All tested flow–stratified discrete sampling schemes were unable to statistically preserve the true load. Composite sampling using a flow–stratified approach provided no statistical advantage or disadvantage over flow discrete sampling but did offer an economic advantage in that fewer samples could be analyzed and still maintain the same absolute error. When a nearly equal number of samples was used, the flow–stratified approach provided a marked decline in absolute error.

Prior to implementation of a sampling program, it is recommended that some knowledge of the watershed runoff characteristics be obtained and studied, if possible. A thorough understanding of the monitoring goal (e.g., will loadings need to be calculated or are concentrations sufficient?) is also important. Expectant length of runoff should also be considered. Short–duration runoff events should not have long times or large flows between samples. Conversely, a long–duration runoff event should not use too small a sampling time or flow to avoid a large number of samples. Water quality planners and managers informed with some knowledge of the watershed runoff characteristics coupled with these reported findings should be able to confidently implement a water quality monitoring program capable of representing the true load while meeting economic constraints.

References

Allen, P. B., N. H Welch, E. D. Rhoades, and C. D. Edens. 1976. The modified Chickasha sediment sampler. ARS–S–107. Washington D.C.: USDA–ARS.

Chapman, B., G. Cooke II, and R. Whitehead. 1967. Automated analysis: The determination of ammoniacal nitrous and nitrate nitrogen in river waters, sewage effluents, and trade effluents. *J. Inst. Water Pollution Control* 66: 185–188.

Claridge, G. G. C. 1975. Automated system for collecting water samples in proportion to stream flow rate. *New Zealand J. Science* 18: 289–296.

Dendy, F. E. 1973. Traversing–slot runoff sampler for small watersheds. ARS–S–15. Washington, D.C.: USDA–ARS.

Doty, R. 1970. A portable automatic water sampler. *Water Resources Research* 6: 1787–1788.

Federal Interagency River Basin Committee. 1948. Measurement of the sediment discharge of streams. Report No. 8. A study of methods used in the measurement and analysis of sediment loads in streams. Iowa City, Iowa: Iowa Institute of Hydraulic Research, Hydraulic Laboratory.

_____. 1952. The design of improved types of suspended sediment samplers. Report No. 6. A study of methods used in the measurement and analysis of sediment loads in streams. Iowa City, Iowa: Iowa Institute of Hydraulic Research, Hydraulic Laboratory.

Federal Interagency River Basin Committee on Water Resources. 1962. Determination of fluvial sediment discharge. Report Q. A study of methods used in the measurement and analysis of sediment loads in streams. Minneapolis, Minn.: St. Anthony Falls Hydraulic Laboratory.

_____. 1963. Determination of fluvial sediment discharge. Report No. 14. A study of methods used in the measurement and analysis of sediment loads in streams. Minneapolis, Minn.: St. Anthony Falls Hydraulic Laboratory.

Gilbertson, C. B., T. M. McCalla, J. R. Ellis, O. E. Cross, and W. R. Woods. 1971. Runoff, solid wastes, and nitrate movement on beef feedlots. *J. Water Pollution Control* 43(3): 483–493.

Harrold, L. L., and D. B. Krimgold. 1943. Devices for measuring rates and amounts of runoff employed in soil conservation research. SCS–TP–51. Washington, D.C.: USDA–SCS.

Holt, R. F. 1969. Runoff and sediment as nutrient sources. In *Conference Proc. Minnesota Chapter Soil and Water Conservation Society of America*, 35–38. Minneapolis, Minn.

Keup, L. E. 1968. Phosphorus in flowing waters. *Water Research* 2: 373–386.
Martin, R. P., and R. E. White. 1982. Automatic sampling of stream water during storm events in small remote catchments. *Earth Surface Processes and Landforms* 7(1): 53–61.
Miller, R. B. 1963. Plant nutrients in hard beech: I. The immobilization of nutrients; II. Seasonal variation in leaf composition; III. The cycle of nutrients. *New Zealand J. Science* 6: 365–413.
Miller, P. S., B. A. Engel, and R. H. Mothar. 2000. Sampling theory and mass load estimation from watershed water quality data. ASAE Paper No. 003050. St. Joseph, Mich.: ASAE.
Mutchler, C. K. 1963. Runoff plot design and installations for soil erosion studies. ARS–41–79. Washington, D.C.: USDA–ARS.
Parsons, D. A. 1954. Coshocton–type runoff samplers–laboratory investigations. SCS–TP–124. Washington, D.C.: USDA–SCS.
Rausch, D. L., and J. D. Haden. 1974. Columbia spillway sampler. ARS–NC–16. Washington, D.C.: USDA–ARS.
Richards, R. P., and J. Holloway. 1987. Monte Carlo studies of sampling strategies for estimating tributary loads. *Water Resources Research* 23(10): 1939–1948.
Robertson, D. M., and E. D. Roerish. 1999. Influence of various water quality sampling strategies on load estimates for small streams. *Water Resources Research* 35(12): 3747–3759.
Romkens, M. J. M., D. W. Nelson, and J. V. Mannering. 1973. Nitrogen and phosphorus composition of surface runoff as affected by tillage method. *J. Environmental Quality* 2(2): 292–295.
Schuman, G. E., R. E. Burwell, R. F. Piest, and R. G. Spomer. 1973. Nitrogen losses in surface runoff from agricultural watersheds on Missouri Valley loess. *J. Environmental Quality* 2(2): 299–302.
Shih, G., W. Abtew, and J. Obeysekera. 1994. Accuracy of nutrient runoff load calculations using time–composite sampling. *Trans. ASAE* 37(2): 419–429.
Stevens, R. J., and R. V. Smith. 1978. A comparison of discrete and intensive sampling for measuring the loads of nitrogen and phosphorus in the River Main, County Antrim. *Water Research* 12(10): 823–830.
Tate, K. W., R. A. Dahlgren, M. J. Singer, B. Allen–Diaz, and E. R. Atwill. 1999. Timing, frequency of sampling affect accuracy of water–quality monitoring. *California Agric.* 53(6): 44–48.
Taylor, A. W., W. M. Edwards, and E. C. Simpson. 1971. Nutrients in streams draining woodland and farmland near Coshocton, Ohio. *Water Resources Research* 7: 81–89.
Thomas, R. B., and J. Lewis. 1995. An evaluation of flow–stratified sampling for estimating suspended sediment loads. *J. Hydrology* 170: 27–45.
Yaksich, S. M., and F. H. Verhoff. 1983. Sampling strategy for river pollutant transport. *J. Environmental Eng.* 109: 219–231.