Evaluating stormwater sampling approaches using a dynamic watershed model

Drew Ackerman, Eric D. Stein and Kerry J. Ritter

Abstract

Accurate quantification of stormwater pollutant levels is essential for estimating discharges to receiving waters as required by many monitoring programs. Numerous sampling approaches exist that attempt to balance accuracy against level of effort (i.e., cost) required to collect the samples. This study employs a novel approach by evaluating the accuracy of different stormwater monitoring methodologies via the output from a continuous simulation watershed model. Seventy eight distinct methodologies were evaluated by "virtual sampling" of fourteen years of model output for Ballona Creek near Los Angeles, California. The 78 methods can be grouped into four general sampling strategies (with numerous permutations): volume-paced compositing, time-paced compositing, pollutograph sampling, and microsampling. The performance of each sampling strategy was evaluated by comparing the median relative error (bias) between the virtually sampled and the true modeled event mean concentration (EMC) of each storm. As a combined measure of bias and precision, the percentage of storms where sampling methods yielded estimates within acceptable levels of accuracy (i.e., 10% of true EMC) were computed across various categories of storm sizes. Finally, costs associated with site setup, personnel costs while sampling and laboratory costs were estimated for each of the methods. Pollutograph sampling consistently outperformed the other three methods both in terms of bias and accuracy. However, pollutograph sampling was the most costly method evaluated. Time-paced sampling consistently underestimated while volume-paced sampling over estimated the storm EMCs. Microsampling performance approached that of pollutograph sampling at a substantial cost savings. The most efficient method for routine stormwater monitoring in terms of a balance between performance and cost was volume-paced composite sampling, with variable sample pacing to ensure that the entirety of the storm was captured. Pollutograph sampling is recommended if the data are to be used for detailed analysis of runoff dynamics.

Introduction

Accurate measurement of stormwater pollutant concentrations is essential for assessing regulatory discharge permit compliance and evaluating the effectiveness of management actions. Decisions based on these data may have human health, ecological and cost implications in terms of development of total maximum daily loads (TMDLs), installation and maintenance of stormwater best management practices (BMPs), or design and implementation of environmental restoration projects.

The inherent variability of stormwater makes consistently accurate measurements of pollutant concentrations challenging. Constituent concentrations can vary within and between storms as a function of watershed characteristics, rainfall, and antecedent dry conditions (Tiefenthaler *et al*. 2001, King *et al*. 2005, Stein *et al*. 2005). Strategies used to account for this variability differ based on sampling objectives, regulatory requirements, cost and logistic considerations. However, most programs' stated goal is to collect the most accurate data possible given technical and resource constraints.

Previous studies have evaluated the effect of various sampling strategies on estimates of pollutant concentration and load (Izuno *et al*. 1998, Robertson and Roerish 1999, Stone *et al*. 2000, Ma *et al*. 2009). Leecaster *et al*. (2002) evaluated sampling approaches based on over 1,700 total suspended solid samples collected at 15-minute intervals from the Santa Ana River, California. They determined that a volume-interval sampling design is the most effective method, and a volume-weighted estimator is best for determining constituent event mean concentrations. Furthermore, Leecaster *et al*. (2002) found that 12 samples per storm were preferable over 4 or 8 samples to avoid excessive variability in estimation results. Ma *et al*. (2009) came to a similar conclusion based on a statistical simulation of various sampling strategies to estimate EMC of chemical oxygen demand (COD). They concluded that volume-paced sampling was superior to time-paced sampling and that approximately 20 samples are required to estimate the EMCs within 20% error for volume-paced sampling. Stone *et al*. (2000) compared four methods for calculating water quality from a 2,050-ha watershed located in North Carolina. Their study concluded that volume-paced sampling provided significantly different and more accurate loads of nitrate- N, ammonia-N, and total Kjeldahl nitrogen when compared with time-paced composite and grab sampling techniques. Other studies have attempted to evaluate a broader set of conditions by using available flow data from numerous stream gauges over multiple watersheds. King and Harmel (2003) evaluated 45 commonly used sampling strategies using 300 storm hydrographs from 87 different watersheds in the United States and provided a set of recommendations based on the goals of specific sampling programs. In general, King and Harmel (2003) concluded that volume-stratified sampling had less absolute error than time-based approaches. However, a major limitation of their study was a lack of concentration data, which required them to base their analysis on hypothetical time vs. concentration relationships ranging from 100% positive to 100% negative correlation with flow.

A limitation of most previous studies is the difficulty in compiling concentration data over the duration of storms for a variety of watershed and climatic conditions. Previous approaches include analyzing data from multiple independent watersheds (Robertson and Roerish 1999) using numerical approaches (Shih *et al*. 1994), Monte Carlo simulations (Richards and Holloway 1987), or statistical approaches (King and Harmel 2004, King *et al*. 2005, Ma *et al*. 2009).

Watershed models provide a previously unutilized, yet innovative tool for comparing the ability of various stormwater sampling approaches to estimate constituent concentrations and loads. Models can simulate a large number of storms for characterizing multiple sampling strategies over a broad range of conditions in a given watershed. In this way, the effect of watershed characteristics noted by King *et al*. (2005) and others can be eliminated. Model output can be used to produce numerous "virtual" sampling schemes and compare them to "true" event mean concentrations and loads based on model output at very short time-steps.

The goals of this study were to use a calibrated and validated watershed model developed for southern California to evaluate multiple stormwater sampling approaches to answer the following questions: 1) What is the relative bias of estimates associated with various stormwater sampling strategies? 2) What percentage of storms for each sampling strategy provide estimates falling within acceptable levels of accuracy? 3) How does bias and percent of storms meeting pre-defined accuracy criterion vary as a function of storm size? 4) What is the tradeoff between performance and cost among the various stormwater sampling strategies?

METHODS

Seventy eight distinct stormwater sampling approaches were evaluated using output from a validated watershed model, which was used to generate 166 storms over a ten year simulation. For each simulated storm, an EMC was calculated for total suspended solids (TSS), fecal coliform and total copper using five-minute model output for the entire storm duration. This was considered the "true value" for the purpose of our analysis as it represents a comprehensive and highly temporally resolved estimate of the EMC. The true value was then compared to a series of EMCs generated by subsampling the model output to represent each of the 78 sampling strategies being evaluated. In all cases, storm EMCs were calculated as:

$$\mathrm{EMC} = \frac{\sum_i V_i C_i}{\sum V}$$

where V_i is the volume that flowed past between sampling times i and i+1, C_i is the sample concentration at time i, and V is the total sampled volume.

The performance of each method was evaluated based on its bias and the percentage of storms meeting acceptable levels of accuracy, relative to the true EMC. Cost were estimated for each of the 78 strategies and compared to the various performance measures to support a "benefit-cost" comparison.

Setting

The 330-km^2 Ballona Creek watershed in the greater Los Angeles, CA, area is an excellent representative urban watershed for analysis. Approximately 85% of the watershed is developed and receives urban runoff. The watershed averages 20 storms and 34 cm (13.4 in) of precipitation per year (Ackerman and Weisberg 2003). Portions of the upper watershed receive as much as 53 cm of percipitation annually, mostly due to orographic effects on south facing slopes of the coastal foothills (Daly and Taylor 1998). Seventy percent of the annual rainfall occurs between January and March, with virtually no rain from May through October (Ackerman and Weisberg 2003). There are no significant point source discharges, stormwater diversion, or detention facilities in the watershed.

Model Calibration and Validation

The analysis relied on a previously calibrated and validated hydrologic and water quality model developed for the Ballona Creek watershed and the nearby, less developed, Malibu Creek watershed (Ackerman *et al*. 2005, Ackerman and Weisberg 2006). The Hydrological Simulation Program—FORTRAN (HSPF; Bicknell *et al*. 2001) was used, which predicts flow based on rainfall, land cover, and stream geometry. Hourly rainfall data for the Ballona Creek watershed was obtained from Los Angeles County Department of Public Works (LACDPW) Gage 10A, Los Angeles International Airport (LAX), and the University of Southern California (USC; Los Angeles County Department of Public Works 2003, NCDC 2004). All gages measured rainfall in 0.25 mm (0.01 in) increments. Land use data were obtained from the Southern California Association of Governments (SCAG 2004; Table 1).

The water quality component of HSPF was calibrated against runoff water quality data collected at sub-hourly intervals from 24 homogenous land use sites distributed across the greater Los Angeles area over a 5-year period from 2001 to 2006 (Tiefenthaler *et al*. 2008). The homogenous land use catchments ranged from 0.02 to 9.49 km^2. Approximately 10 samples were collected per storm event at 30- to 60-minute intervals for each site-event depending on the size and timing of the storm. Samples were collected more frequently when flow rates were high or rapidly changing and less frequently during lower flow periods. This allowed for characterization of concentration changes over the course of each storm and incorporation of this temporal resolution into the model calibration. Calibration samples were analyzed for TSS, metals, nutrients, and bacteria. The calibrated model was then applied at the watershed scale in Ballona Creek and validated using data from seven storms collected at the bottom of the watershed in the same manner (and analyzed for the same constituents) as that collected from the land use sites. Predicted EMCs from the Ballona Creek model for TSS, total copper, and fecal coliform were within 28, 8, and 12% of measured values, respectively (Ackerman and Weisberg unpublished data).

Table 1. Land use area and Rational Method runoff coefficient for the Ballona Creek watershed model. Note that Transportation Land Use is included as part of the other land use categories and not as an independent category.

Land Use	Area (km^2)	Runoff Coefficient
Agriculture	0.09	0.10
Commercial	39.36	0.61
Industrial	15.11	0.51
Residential	154.31	0.39
Mixed Urban	0.40	0.41
Open	38.65	0.06

Application of Model to Evaluate Sampling Strategies

The model was used to simulate all storms between water years 1990 and 2004. This fourteen year period contained four years in each of the upper and lower quartiles of annual rainfall (based on data from 1948 to 2004) making it a good representative time period for analysis. There were 166 sampleable storms in the fourteen year simulation with sizes ranging from 0.25 to 13.1 cm (0.1 to 5.17 in). Sampleable storms in the model output were defined based on criteria typically used in stormwater sampling. For a storm to be sampleable, it must have been associated with rainfall events of 0.24 cm (0.1 in) or greater and have occurred following at least three antecedent days with no measurable rain. These criteria reflect those used to trigger compliance sampling under most municipal stormwater permits. The beginning of a storm was defined as a 30% increase in baseflow, based on review of historical flow data for Ballona Creek. The end of a storm was defined in one of two ways: 1) flow increased due to a subsequent rainfall event, or 2) flow receded to a set fraction of the storm peak flow.

The point in the storm where the simulated sampling was terminated has the potential to affect the accuracy of the EMC. Therefore, we evaluated the effect of different "end of storm criteria" for each sampling strategy by calculating the median error between the true storm EMCs based on model output for the entire storm and the EMC resulting from a subset of the model output based on termination of sampling at points ranging from 10 to 99% below the peak flow.

Stormwater Sampling Methods Evaluated

The 78 individual sampling methods analyzed can be categorized into five general strategies; volume-paced composite sampling, time-paced composite sampling, pollutograph sampling, volume-paced microsampling, and time-paced microsampling (Table 2). For each sampling method the five-minute model output was subsampled in a manner that replicates the actual field-based sampling. The EMCs resulting from each virtual sampling was compared to the modeled storm EMC based on the entire five-minute output to assess performance (see Statistical Analysis). Results were compared across all storm sizes as well for small (<6.4 mm; <0.25 in), medium (6.4 to 28mm; 0.25 to1.1 in) and large (>28 mm; >1.1 in) storms. The thresholds between the storm size classes were based on the distribution of storms over the 14-year modeling period.

Volume-paced composite sampling

Volume-paced sampling was designed to mimic the sampling strategy used by many stormwater agencies. In practice, the volume-paced method uses an automated flow meter and sampler. When flow at a given site reaches a specified level above baseline during a storm (here it was 30% above baseflow), the automated sampler is triggered to draw a sample (an aliquot) at pre-set runoff volume intervals. The samples are composited into a single sample jar and analyzed to produce an EMC for that storm.

To replicate the volume-paced sampling in Ballona Creek, past sampling efforts were reviewed and a series of assumptions made about the sampling. Volume pacing in Ballona Creek was set to draw an aliquot at intervals ranging from every 1,400 to 120,000 m^3 (50,000 to 4,250,000 ft^3; Table 2) with a total of 48 aliquots (Table 3). Volume-paced sampling of the model output was defined to mimic the actual field sampling. Based on field experience, we assumed that a maximum of two samples could be drawn in five minutes which incorporated time to draw the sample and back-purge the line. A minimum of 24 samples (half of the targeted) were required to provide adequate volume for analysis. Any storm without 24 samples was assumed to have insufficient sample for laboratory analysis.

The accuracy of each volume pacing was evaluated on a storm-by-storm basis as a function of a targeted storm volume. That is, how does the relationship between actual storm size and the anticipated storm size used to set the interval between collecting successive samples affect accuracy (Table 3). In an operational, field sense, a targeted storm volume is the total runoff of an anticipated storm. Targeted storm volumes were calculated by using a simple model of rainfall and land use via the Rational Method (Ackerman and Schiff 2003; Table 1):

Volume = A x i x c

where: A = Land use drainage area (km^2); i = Rainfall (mm); and c = Runoff coefficient (unitless).

Time-paced composite sampling

An alternative to volume-paced sampling is time-paced sampling, where an aliquot is collected based on equally spaced time increments. The onset of sampling and sample collection is the same as detailed in the volume-paced sampling, but with time setting the pacing of the sample collection. Time-paced sampling was set to capture a range of storm durations. Time-paced samples ranged from every 5 minutes to 1 every hour with a total of 48 samples (Table 2). The pacings corresponded to a total sampling duration ranging from 4 to 48 hours.

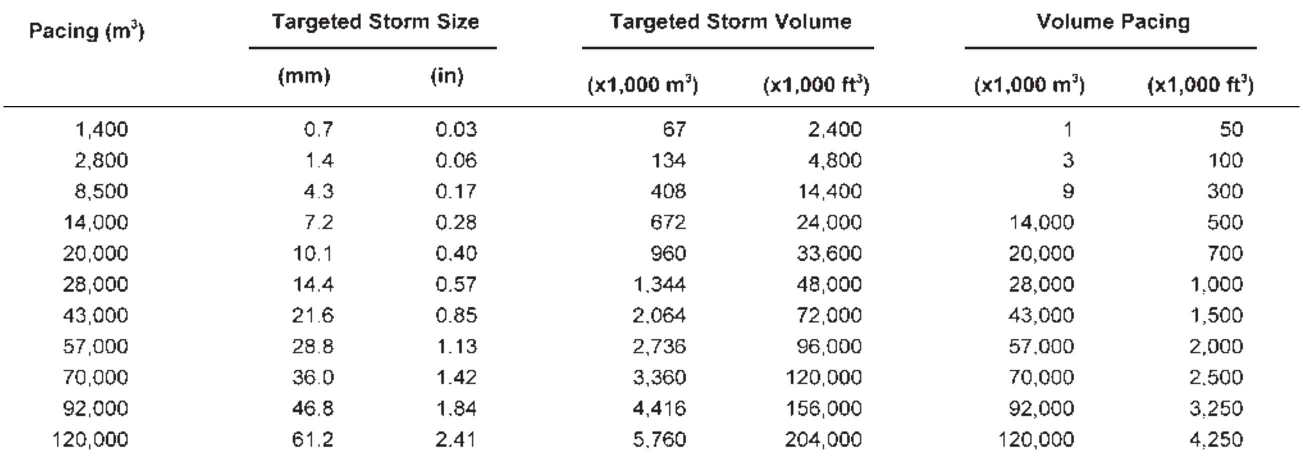
Pollutograph sampling

Pollutograph sampling is a more intensive sampling approach than either the volume- or time-paced sampling, and consists of collecting multiple discrete samples throughout a storm. In practice, a large number (often 20) of samples are collected throughout the storm hydrograph. When storm flows return to near base flow and field sampling discontinues; a subset of the larger number of samples are selected for analysis based on the portion of the hydrograph that is targeted for sampling.

Table 2. Sample pacing for the different virtual methodologies simulated from the model output. Pollutograph and volume-paced microsampling are defined as a percent of the total sampleable storm volume.

Sampling Strategy	Sample Pacing (m^3)	Sample Pacing (ft^3)	Method ID
Volume-paced Composite Sampling	1,400	50,000	Volume_1
	2,800	100,000	Volume_2
	8,500	300,000	Volume_3
	14,000	500,000	Volume_4
	20,000	700,000	Volume_5
	28,000	1,000,000	Volume_6
	43,000	1,500,000	Volume_7
	57,000	2,000,000	Volume_8
	70,000	2,500,000	Volume_9
	92,000	3,250,000	Volume_10
	120,000	4,250,000	Volume_11
	Sample Pacing (minutes)	**Duration (hours)**	
Time-paced Composite Sampling	5	4	Time 1
	10	8	Time 2
	15	12	Time 3
	20	16	Time 4
	30	24	Time 5
	60	48	Time 6
		Sample Pacing (% total storm volume)	
Pollutograph Sampling (4 bottles)		10 30 50 80	Poll 4.1
		20 40 60 80	Poll 4.2
		10 20 40 80	Poll 4.3
		20 60 80 90	Poll 4.4
Pollutograph Sampling (10 bottles)		5 10 15 20 25 30 40 50 70 90	Poll 10.1
		9 18 27 36 45 54 63 72 81 90	Poll 10.2
		5 10 20 30 40 50 60 70 80 90	Poll 10.3
		10 30 50 60 70 75 80 85 90 95	Poll 10.4
		Sample Pacing (% target storm volume)	
Volume-paced Microsampling (2 bottles)		20 60	Micro Vol 1
		30 70	Micro Vol 2
		40 80	Micro Vol 3
Volume-paced Microsampling (4 bottles)		10 20 40 80	Micro Vol 4
		20 40 60 80	Micro Vol 5
		20 60 80 90	Micro Vol 6
Volume-paced Microsampling (8 bottles)		5 10 15 25 35 45 60 80	Micro Vol 7
		11 22 33 44 55 66 77 88	Micro Vol 8
		20 40 55 65 75 85 90 95	Micro Vol 9
	Sample Pacing (minutes)	**Duration (hours)**	
Time-paced Microsampling (2 bottles)	10 20	5	Micro Time 1
	15 30	7.5	Micro Time 2
	20 40	10	Micro Time 3
	30 60	15	Micro Time 4
Time-paced Microsampling (4 bottles)	5 10 20 40	12.5	Micro Time 5
	10 20 40 60	21.7	Micro Time 6
	15 30 60 60	27.5	Micro Time 7
Time-paced Microsampling (8 bottles)	5 10 15 20 30 30 30 30	28	Micro Time 8
	5 10 20 40 40 40 40 40	39.2	Micro Time 9
	10 20 40 60 60 60 60 60	61.7	Micro Time 10

Table 3. Targeted storm volumes for the different volumetric pacings that were evaluated.



Pacing (m^3)	Targeted Storm Size		Targeted Storm Volume		Volume Pacing	
	(mm)	(in)	(x1,000 m^3)	(x1,000 ft^3)	(x1,000 m^3)	(x1,000 ft^3)
1,400	0.7	0.03	67	2,400	1	50
2,800	1.4	0.06	134	4,800	3	100
8,500	4.3	0.17	408	14,400	9	300
14,000	7.2	0.28	672	24,000	14,000	500
20,000	10.1	0.40	960	33,600	20,000	700
28,000	14.4	0.57	1,344	48,000	28,000	1,000
43,000	21.6	0.85	2,064	72,000	43,000	1,500
57,000	28.8	1.13	2,736	96,000	57,000	2,000
70,000	36.0	1.42	3,360	120,000	70,000	2,500
92,000	46.8	1.84	4,416	156,000	92,000	3,250
120,000	61.2	2.41	5,760	204,000	120,000	4,250

Eight different pollutograph strategies were simulated. Pollutograph sampling was evaluated with four different pacings (i.e., the interval at which a sample is collected) and for collection of either four or ten discrete samples. The various pollutograph sampling strategies weighted the first part of the storm to capture higher concentration typically observed early in the hydrograph. Because field crews can often have a time limitation on their sampling, a 24-hour cutoff was evaluated relative to values which would result from sampling the entire storm.

Volume-paced microsampling

Volume-paced microsampling is a combination of volume-paced and pollutograph sampling. With microsampling, numerous small aliquots are drawn at a given volume pacing, depending on the storm size targeted, and composited into between one and eight bottles. For this analysis, we assumed that each bottle was comprised of ten microsample aliquots. Unlike pollutograph sampling, which must be done manually, microsampling allows the use of autosamplers yet still allows samples to be collected across the entire hydrograph for calculation of a volume-paced EMC. We assumed back-purging was not critical in the microsampling, and as such one aliquot could be collected each minute.

Several volume-pacing strategies were evaluated as well as strategies that involved compositing samples into one, two, four, or eight bottles (Table 2). These permutations were targeted based on field experience with autosampler configuration. For example one, two, or four samples could be collected in a single autosampler while eight samples would require two autosamplers to be used in series.

Volume-paced microsampling was evaluated based on targeted storms volumes corresponding to the 25th, 50th, 75th and 90th percentile of all sampleable storms for the 14-year modeling period.

Time-paced microsampling

Time-paced microsampling is analogous to volume-paced microsampling, except that aliquots are drawn at pre-defined time intervals instead of at volume intervals. A variety of permutations were evaluated for storm durations from 5 to 61 hours. Pacings between aliquots varied from 10 to 60 minutes, with the interval between microsamples typically occurring more frequently during the early portions of the storm (Table 2). We assumed that each bottle was comprised of 10 aliquots, as in the volume-paced microsampling strategy. Strategies evaluated included sampling into one, two, four, or eight bottles, as was done for the volume-paced microsampling evaluation.

Threshold for Inclusion of Storms in Analysis

For each method analyzed we included a minimum threshold of storm volume that needed to be captured by the simulated autosampling in order to produce a storm EMC. Below this level, insufficient storm was sampled to calculate a meaningful EMC. For the volume and time-paced composite sampling, at least half of the number of samples targeted by the method needed to be collected. The pollutograph sampling is done manually, which allows sufficient samples to be collected for every storm; hence the threshold analysis is not relevant. Both volume and time weighted microsampling assumed ten aliquots would be collected for each individual sample; therefore, if 50% of the target (i.e., 5 aliquots) could not be collected, the sample was considered not usable.

Statistical Analysis

Evaluating performance among sampling methods across storms relied on robust alternatives to the usual measures of bias. Robust measures were used because distributions of errors were often truncated or skewed and one or two extreme values were common. Therefore, measures based on medians rather than means were a better representation the central tendencies of errors. For example, the "mean" in the relative bias was replaced by the "median" to give the median relative error (MRE),

$$\mathrm{MRE} = med_i\left(\frac{x_i - \theta}{\theta}\right)\times 100\%$$

where: x_i is the parameter estimate for the *ith* storm and θ is the true value of the parameter being estimated.

As a combined performance measure of both the bias and precision inherent among the different sampling methods, we tallied the percentage of evaluated storms within each storm size category that were within acceptable error ranges (within 10% of target EMC value). This percentage is consistent with the threshold recommend by Harmel *et al*. (2003).

All performance measures were computed across all sampled storms within each storm size category (small, medium, and large). Finally, we computed the median percent of the true volume captured by each sampling method storms as a possible correlate to error.

Costs

Sampling decisions are often affected by cost considerations. Therefore, we estimated the cost associated with each sampling strategy relative to its associated level of performance. Costs were based on typical labor and laboratory analytical expenses for the Los Angeles area (Table 4). Estimated costs for each sampling method were conservative because of multiple assumptions. Cost estimates include an assumption that each station would need to be set up prior to sampling; in reality these costs would not occur if an existing station were to be used. Personnel costs include equipment rental, installation, maintenance of the sampling site, personnel training and data management. We also assumed that all of the bottles collected would be analyzed. This is an overestimation at times because some of the microsampling strategies either had insufficient volume to analyze, or not all of the targeted bottles were analyzed. To estimate the storm length of the pollutograph samplings exceeding 24 hours (for labor costs), we averaged the duration of the ten modeled storms that were longer than 24 hours (40 hours average).

RESULTS

Ability of Methods to Adequately Sample a Range of Storms

Pollutograph sampling was successfully applied to nearly all storms regardless of size. The volume and time-paced microsampling approached the performance of the pollutograph sampling (Figure 1), with both volume and time pacings consistently characterized more than 95% of the storm EMCs. The ability of volume-paced composite sampling to adequately capture storms (at least half of the 48 aliquots taken) varied based on the sample pacing used. Only storms with pacings at or below 14,000 m^3 (500,000 ft^3) were able to characterize more than half of the small storms. For medium storms greater than 50% of the storms were captured for pacings less than 57,000 m^3 (2,000,000 ft^3; Figure 1). For time-paced composite sampling, less than half of the small and medium storms were captured when pacings exceeded five minutes. Of the 6 time pacings tested, a 15-minute time pacing resulted in the highest percentage storm capture for all 3 storm sizes.

Effect of "End of Storm" Definition

Stormwater sampling typically extends at least through the peak flow. The decision of how far beyond peak flow to continue sampling affects the error in the EMC. For TSS, copper and bacteria continuing sampling until the flow reached 50% of peak flow resulted in approximately 10% median error (between EMC defined by the peak flow termination criteria and the true EMC). The one exception was for bacteria during large storms, where the error at 50% of peak flow was greater than 15%. Extending sampling until flow was 75% below peak flow reduced the median error to around 5%; beyond that the incremental decreases in error were small.

Table 4. Costs associated with each sampling strategy. Assumptions include: the site is set-up; the equipment is installed and programmed; travel time is less than 2 hours; all permits and site access are complete; siting and design is complete; extended sampling at rates are $200/hr/team and general chemistry costs are $800/bottle. Pollutograph sampling assumes staff remain on-site to conduct manual collection of samples.

Sample Type	Site Setup	Personnel	Time (hour)	Bottles	Chemistry Costs	Total Cost
Volume-paced Composite	$1,500	$4,000	-	1	$800	$6,300
Time-paced Composite	$1,000	$4,000	-	1	$800	$5,800
Pollutograph (24-hour cutoff)	$2,600	$6,000	24	4	$3,200	$11,800
Pollutograph (24-hour cutoff)	$2,600	$6,000	24	10	$8.000	$16,600
Pollutograph (no cutoff)	$2,600	$10,000	40	4	$3.200	$15,800
Pollutograph (no cutoff)	$2,600	$10,000	40	10	$8.000	$20,600
Volume-paced Microsampling	$2,500	$4,000	-	2	$1.600	$8,100
Volume-paced Microsampling	$2,500	$4,000	-	4	$3.200	$9,700
Volume-paced Microsampling	$2,500	$4,000		8	$6.400	$12,900
Time-paced Microsampling	$2,500	$4,000	-	2	$1.600	$8,100
Time-paced Microsampling	$2,500	$4,000	-	4	$3,200	$9,700
Time-paced Microsampling	$2,500	$4,000	-	8	$6,400	$12,900

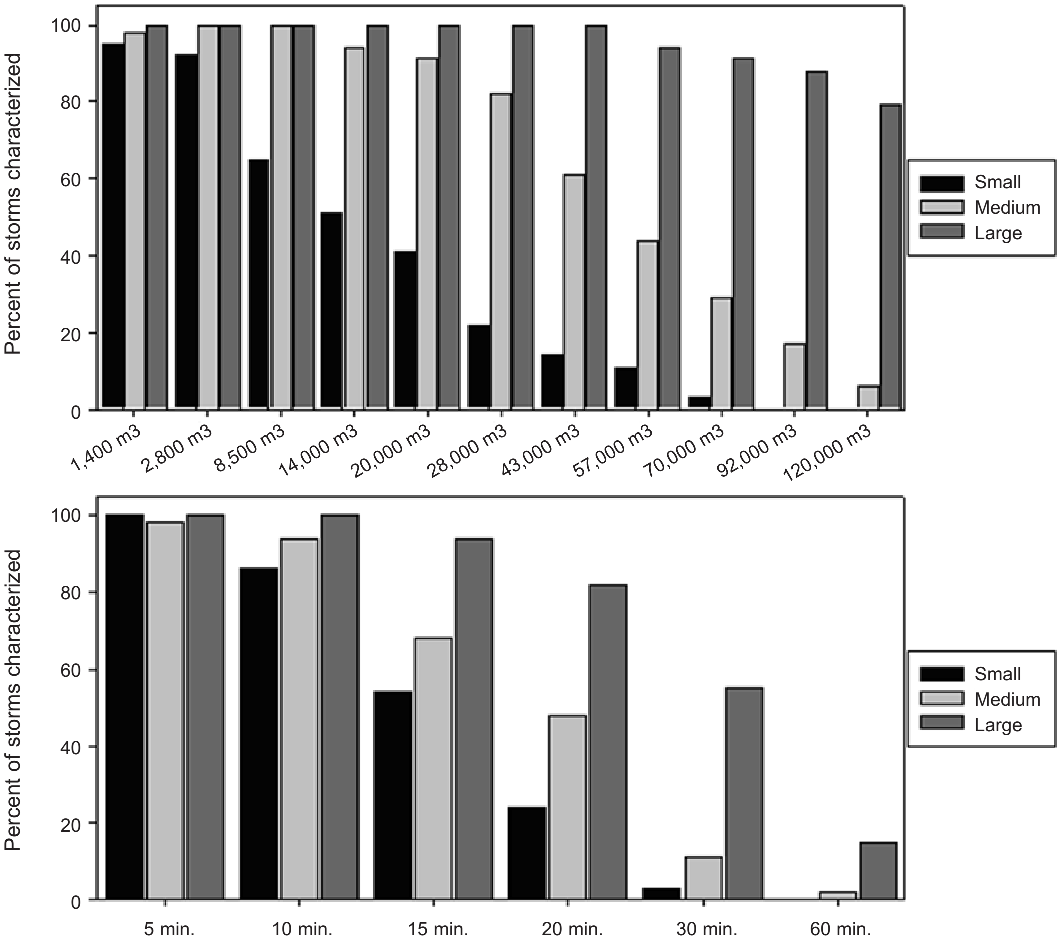


Figure 1. Percent of total storms that were sampleable for various permutations of the volume-paced (top) and time-paced samplings (bottom). Small = <6.4 mm; Medium = 6.4 - 28 mm; Large = >28 mm.

Sampling Bias

Pollutograph and volume-paced microsampling produced estimates with the smallest bias (as measured by MRE) among all sampling methods. In both cases the MRE was within +10% for most pacing variations, with bias being slightly higher for larger storms than for smaller storms (Figure 2). There was no significant increase in bias when a 24-hour cutoff was imposed on the sampling. Time-paced microsampling for small and medium storms had slightly negative bias for copper and slightly positive bias for bacteria. Volume-paced composite sampling resulted in the largest overestimates of EMC, particularly for large storms. As the pacing increased, and thus the target storm size, to the break in storm size definitions, the magnitude of bias decreased for both constituents, becoming negative as the pacing increased past target volume. Time-paced composite sampling resulted in the largest negative bias with MRE values up to -20%. Bias was substantially higher for each sampling strategy when estimating EMCs for fecal coliform than for copper (and TSS; Figure 2), with bias up to +50% for small and medium storms and up to +150% for large storms.

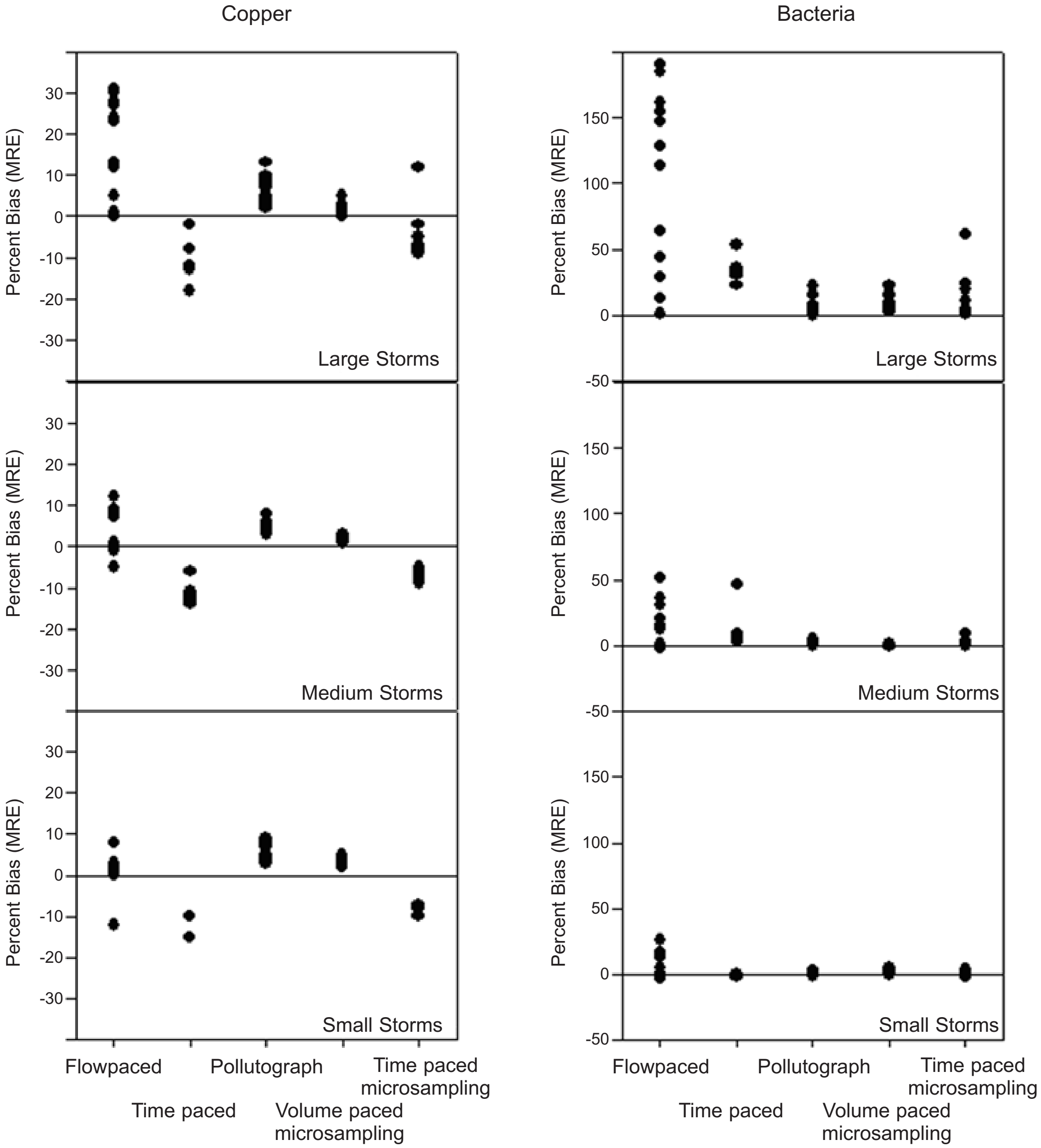


Figure 2. Bias of each general method grouping for copper and bacteria for small medium and large storms. Each point represents a distinct permutation within each general group.

Sampling Accuracy

Sampling accuracy was measured by quantifying storms within ten percent of the true EMC for a given strategy, pacing, and storm size (Figure 3). The general patterns for accuracy were similar to those for bias, with pollutograph sampling having the highest accuracy and the volume and time-paced compositing methods having the lowest. Unlike the bias results, there was a greater difference based on pacings within a given sampling strategy and the results varied less as a function of storm size. Also, unlike bias, accuracy results were much more comparable for copper and bacteria sampling.

Pollutograph sampling consistently had higher accuracy than the composite sampling (Figure 3). The 10-bottle pollutograph sampling with evenly distributed bottles (Poll 10.2), with and without a 24-hour cutoff, resulted in more than 97% of copper EMCs within 10% of the true EMC and 70% of the bacteria EMCs. The comparable 4-bottle sampling (Poll 4.2) resulted in greater than 65% of copper EMCs and 76% of the bacteria EMCs within 10% of the true EMC for medium and small storm events.

Volume-paced microsampling resulted in the most consistent accuracy of any method evaluated (i.e., there was the least variability in performance as

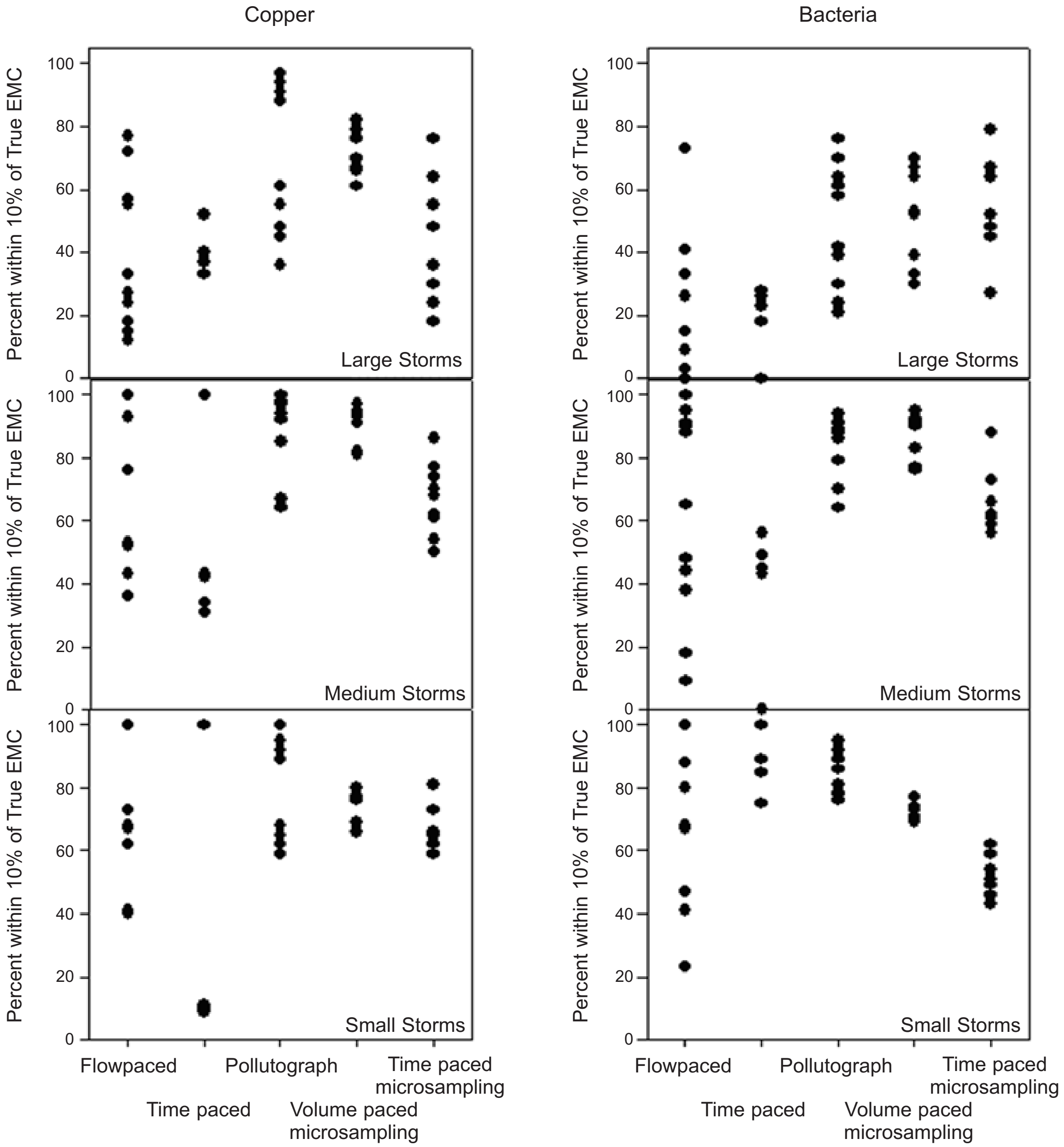


Figure 3. Percent of storms within each general method grouping that were within 10% of the true EMC for small medium and large storms. Each point represents a distinct permutation within each general group.

a function of storm size or sample pacing). Volume-paced microsampling performed best during medium storms with accuracy generally within 10% of the true EMC 90% of the time for both copper and bacteria. Accuracy was slightly less for copper in the large and small storms and considerably less, but comparable to the pollutograph sampling, for the largest storms. Time-paced microsampling was generally less accurate than volume-paced microsampling, with performance decreasing for the largest events.

Volume-based composite sampling resulted in the greatest variability in accuracy as a function of sample pacing. For some of the volume pacings (e.g., pacing greater than 28,000 m^3/1,000,000 ft^3 per sample) greater than 88% of the sampled storms characterize the storm EMC within 10% of the true EMC. However, those high pacings were unable to characterize more than half of the storm events, making their implementation unlikely. As with bias, accuracy was greater for copper than for bacteria, particularly as pacings increased to be close to the target storm volume. The accuracy of time-based composite sampling was substantially lower than that of volume-based composite sampling. In most cases, less than half of the pacing variations for time-based composite sampling were within 10% of the true EMC, with performance being lower for bacteria than for copper.

By using an adaptive, targeted sampling, where volumetric pacings are adjusted based on the anticipated storm size, the performance of volume weighted composite sampling can approach that of pollutograph sampling. Figure 4 shows that if the volume of an incoming storm is correctly estimated, nearly 100% of the samplings would have their EMCs within 10% of the true EMC. If the storm volume is lower than expected, the performance is equally high. If, however, the storm volume is higher than expected, performance decreases to nearly 40% of samples within 10% of the true EMC when the actual storm volume is nearly double the anticipated (and hence targeted) volume.

Bias and Accuracy vs. Cost

Sampling strategies that resulted in the lowest bias and the highest accuracy (i.e., pollutograph sampling) were the most expensive ($11,000 - $20,000 per storm; Figures 5 and 6). However, the cost of pollutograph sampling can be reduced by reducing the number of bottles from ten to four, with a moderate reduction in performance. Ending sampling after 24 hours resulted in cost savings of up to $4,800 per storm, with little decrease in performance. Volume and time based composite sampling had the lowest costs ($5,800 to $6,300), but also the lowest accuracy and highest bias. Costs for microsampling strategies were intermediate between pollutograph and composite sampling ($8,000 to $12,000 per storm). Similarly, bias and accuracy were close to, but slightly poorer than those obtained for pollutograph sampling.

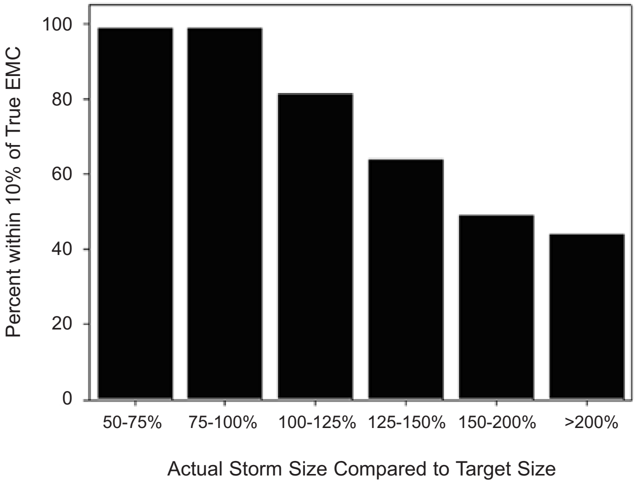


Figure 4. Accuracy of targeted volume-paced sampling for various storm size targeting.

DISCUSSION

The goal of stormwater sampling is to obtain an accurate representation of the concentrations and loads of pollutants of concern being discharged from catchments or watersheds of interest. The dynamic and somewhat unpredictable nature of storms makes it challenging to obtain accurate estimates of concentration in a cost-effective manner. Because the results of stormwater sampling are often used to assess regulatory compliance and/or make management decisions that affect human and ecological health, there is a strong need to maximize the accuracy of estimates. The results of this study show that the choice of sampling strategy affects how well EMC measurements estimate the actual concentrations being discharged. The 10-bottle pollutograph strategy produces the most accurate EMC estimates and has the least bias. Because pollutograph sampling involves having people in the field for the duration of the storm, it also affords the greatest flexibility and ability to adapt to changing conditions over the course of a storm. The time-variable con-

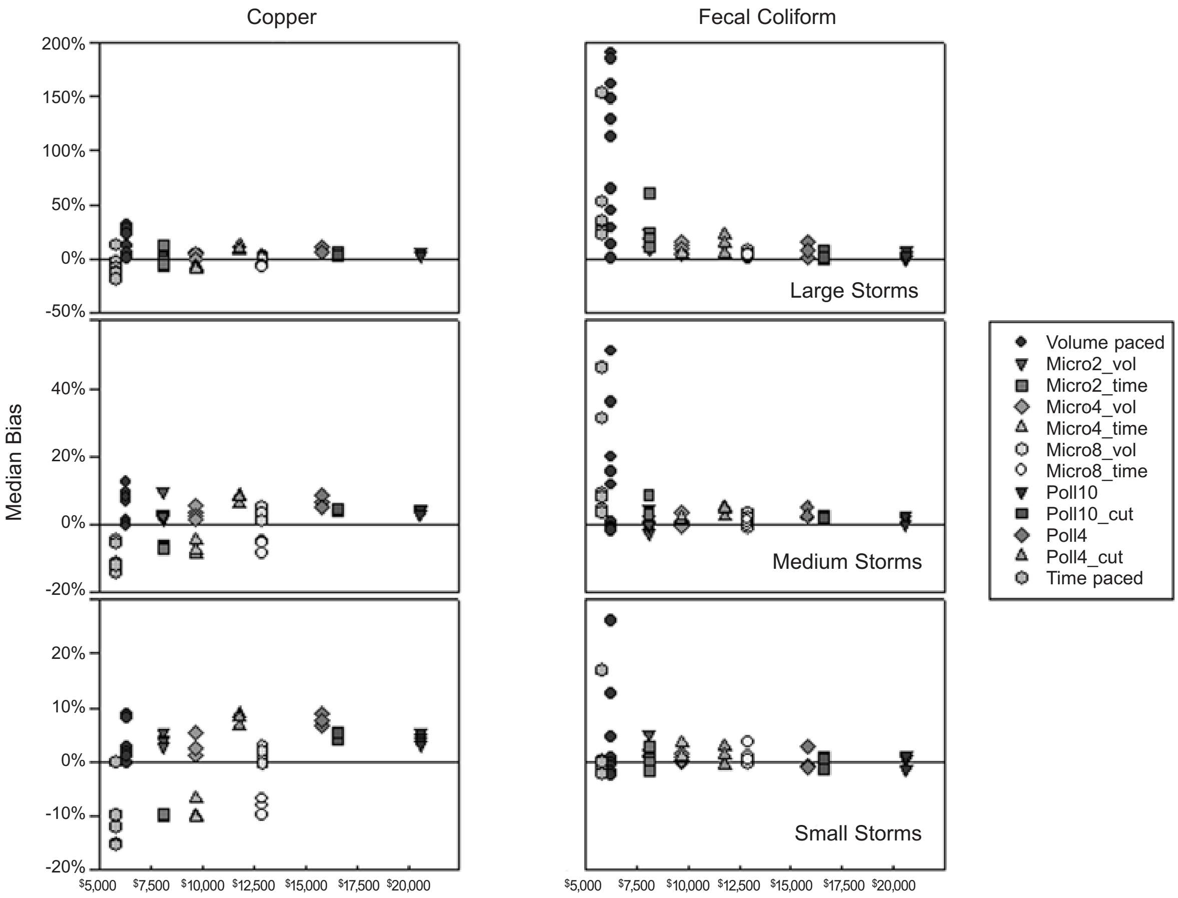


Figure 5. Copper and bacteria bias as a function of cost by storm size. Only storms with greater than half of the captured storms are shown. Each point represents a different sampling pacing for a given method.

centrations obtained by pollutograph sampling are useful and sometimes necessary for detailed characterization of runoff patterns or for model calibration (Stein *et al*. 2005, Tiefenthaler *et al*. 2008). However, pollutograph sampling is the most costly approach and may not be warranted for stormwater sampling associated with routine monitoring or regulatory compliance.

The most commonly applied (and least costly) sampling strategies are time and volume-paced composite sampling. These approaches are the least accurate and have the highest bias. Consequently, they may result in erroneous results that lead to inappropriate conclusions regarding concentrations and annual loads being discharged to receiving waters.

Volume-paced microsampling and targeted volume-paced sampling with analysis of discrete samples provide alternatives that improve accuracy without costing as much as pollutograph sampling (Figure 7). The common features of both these approaches are: 1) use of volume pacing, not time pacings, 2) their ability to capture a range of different storm types (i.e., sizes and timing), and 3) their inclusion of multiple discrete samples. Numerous authors have previously documented that volume-based sampling is more accurate than time-based because it provides better representation of the overall storm (Leecaster *et al*. 2002, King *et al*. 2005, Ma *et al*. 2009). By targeting the volumetric pacings based on anticipated storm size, sampling is better able to capture a representative portion of the storm. Given the error inherent in weather predictions, it is preferable to overestimate (i.e., storm is smaller than expected) than to underestimate when setting the sample pacings (Figure 4). Although more costly, analyzing discrete samples as opposed to compositing into a single sample allows for better representation of changing concentrations over the course of a storm, results in more accurate EMCs and provides greatest flexibility if the storm does not materialize as predicted.

Costs can be partially reduced by shortening the duration of sampling or reducing the number of dis-

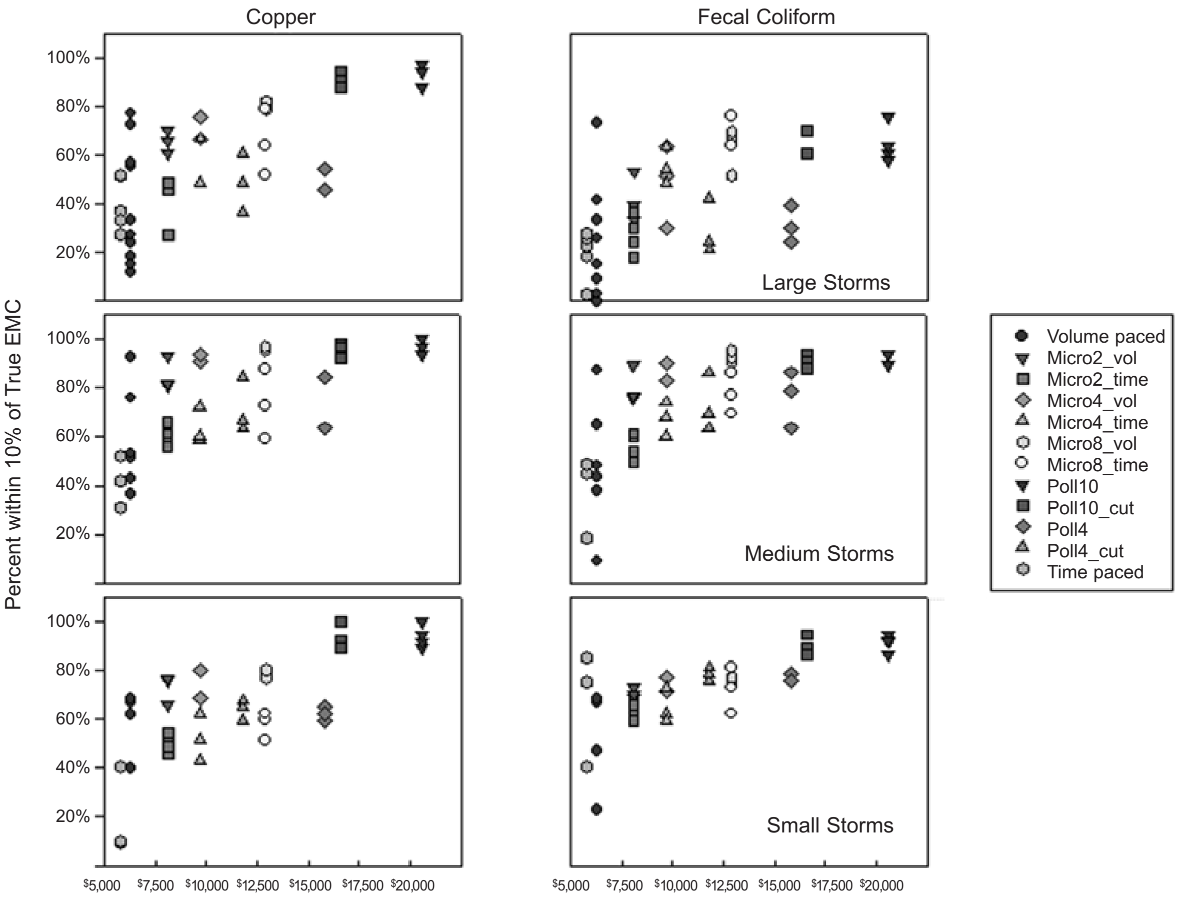


Figure 6. Copper and bacteria EMCs that are within 10% of the true EMC by storm size. Only methods with greater than half of the captured storms are shown. Each point represents a different sampling pacing for a given method.

crete samples analyzed from ten to four. Because concentrations are typically higher during the early portion of storms (Stein *et al*. 2005, Tiefenthaler *et al*. 2008), ending sampling when flow is 50% of peak flow can reduce costs with little overall effect on accuracy of the EMCs because at that point, the majority of the pollutant mass and storm volume have flowed past the station. However, this requires accurate assessment of the timing of peak flow in the field; inaccurate determination of peak flow may introduce additional bias. If the goal of sampling is to produce an overall storm EMC, reducing the number of discrete samples from ten to four results withevenly distributed sampling medium storms, resulted in a decrease in accuracy from 94 to 85%.

Automated bacteria sampling is complicated by the need to transport samples to the lab within six hours. Consequently, bacteria concentrations are often estimated based on a single grab sample. We simulated this approach by randomly selecting a bacteria sample during the first, second, or third hour of each of the 166 storms modeled. The accuracy of this approach varied from less than 10% of samples being within 10% of the actual EMC for large storms to 30% of samples being within 10% of the actual EMC for small storms. The accuracy of EMC estimates can be improved if the grab samples are taken from one of the composited bottles, each containing an aggregate a portion of the storm (assuming the bottles have been pre-sterilized). If the sample is taken relatively early in the storm, the accuracy can improve to up to 50% of samples being within 10% of the actual EMC, although results are better for small storms than for large (Figure 7). However, to obtain accuracies similar to those of TSS and copper, multiple composite samples must be collected and delivered to the lab several times during the course of a storm.

Insights from this study can be used in designing stormwater sampling programs. For example, if a program is designed for compliance and only a single EMC for a storm is desired, great gains can be achieved in both accuracy and bias by targeting a storm volume before sampling begins. If the

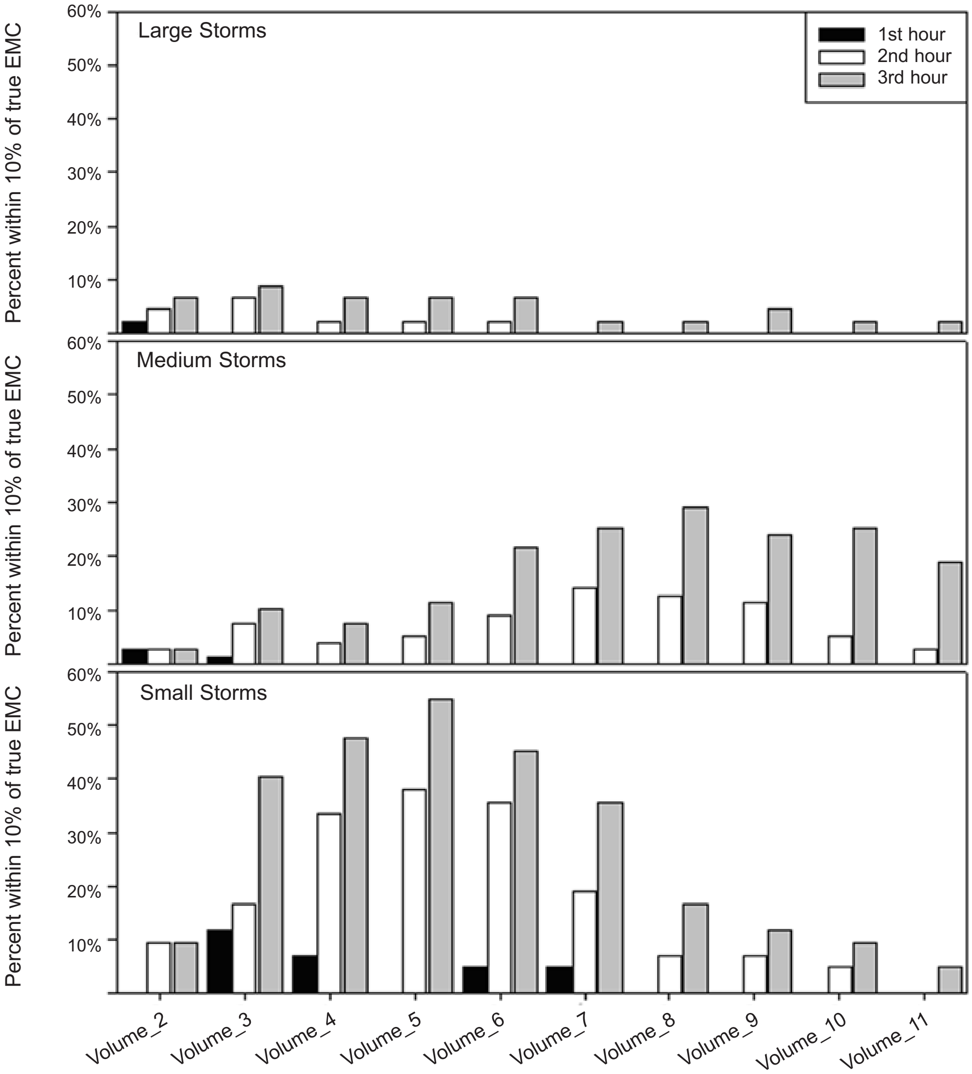


Figure 7. Fecal coliform concentrations within 10% of the true EMC when samples are taken from the volume-paced composite within the first one, two, and three hours. Small = <6.4 mm; Medium = 6.4 - 28 mm; Large = >28 mm.

stormwater data is being collected for a more intensive study where information on pollutant behavior is desired (e.g., modeling, BMP design/testing), using automated microsampling can have considerable cost savings while still obtaining high quality data as would be collected in pollutograph sampling.

Stormwater sampling programs are often developed to comply with regulatory requirements, such as Total Maximum Daily Loads (TMDLs). Those loads are typically calculated by estimating a storm EMC and multiplying that by the total storm volume to obtain a load. The efforts in this paper have intentionally not focused on load because methods to estimate storm volumes and their uncertainties have been addressed previously. Another complication of sampling to support TMDL compliance is that water quality standards, such as the California Toxics Rule (CTR), are based on dissolved concentrations; whereas, stormwater samples are typically analyzed for total concentrations. To relate our analysis to this discrepancy, we assumed a hardness of 100 mg/L, and an associated CTR limit for copper of 13.4 µg/L dissolved. The percent dissolved copper in Ballona Creek stormwater varies between 11 to 60% (Buffleben *et al*. 2002). Using these values, pollutograph and microsampling slightly over-predicted the exceedence frequency and performed much better than the general volume weighted sampling (Table 5) but were consistently within 20% of the modeled EMCs (Figures 3 and 6).

Our results are consistent with, and build upon, the findings of previous studies. The accuracy we found with the 10-bottle pollutograph sampling compares well with the results of Leecaster *et al*. (2002). Our assessment that volume-paced sampling was superior to time-paced echoes the findings of King and Harmel (2004). The microsampling approach introduced in this study has similar advantages to the "extended grab sampling" recommended by Ma *et al*. (2009), but does not require manual collection of samples and allows for greater integration over portions of the storm.

This study builds on previous work by demonstrating how a calibrated and validated watershed model can be used to simulate sampling approaches in a realistic manner for an extended duration (14 years in this study; Ackerman and Weisberg 2006). A comparable empirical analysis would be cost and time prohibitive. The model also allows for evaluation of sampling methods over a variety of storm types and produces a large enough sample size to support robust statistical analysis. Furthermore, the model allows for more direct evaluation of the effect of sampling strategies on the accuracy of EMCs for pollutants of concern.

The results of this study should be applicable for many urban stormwater monitoring and assessment programs. Conclusions may differ in areas where the hydrograph shape differs from the short, flashy storms typical of southern California urban watersheds. For example, sampling strategies for snowmelt runoff dominated storms may have to be modified to accommodate a longer, lower amplitude hydrograph. King and Harmel (2004) noted that volume based sampling was sensitive to watershed parameters such as hydraulic length, slope, curve number, and runoff coefficient. The strength of using a watershed model is that it provides a tool that can be relatively easily adapted to investigate the effect of various sampling strategies for specific watershed types, sizes, or climatic conditions.

Table 5. Percent of samples exceeding CTR copper water quality standards (13.4 µg/L assuming 100 mg/L hardness) for different sampling methods at differing levels of percent dissolved copper.

Sampling Method	Percent Dissolved			
	20	25	30	40
Model	1	19	59	93
Volume-weighted (at 8,500 m^3)	12	57	95	100
Time-weighted (at 10 minutes)	0	4	26	93
4-bottle Pollutograph (evenly weighted, 24-hour cutoff)	4	31	72	94
4-bottle Pollutograph (evenly weighted)	3	26	70	97
4-bottle Microsampling (evenly weighted)	0	28	70	98

Pollutograph sampling consistently outperformed the other three methods both in terms of bias and accuracy. However, pollutograph sampling was the most costly method evaluated. Time-paced sampling consistently underestimated while volume-paced sampling over estimated the storm EMCs. Microsampling performance approached that of pollutograph sampling at a substantial cost savings. The most efficient method for routine stormwater monitoring in terms of a balance between performance and cost was volume-paced composite sampling,with variable sample pacing to ensure that the entirety of the storm was captured. Pollutograph sampling is recommended if the data are to be used for detailed analysis of runoff dynamics, such as is required for model calibration.

LITERATURE CITED

Ackerman, D. and K. Schiff. 2003. Modeling stormwater mass emissions to the Southern California Bight. *Journal of Environmental Engineers* 129:308-317.

Ackerman, D., K. Schiff and S. Weisberg. 2005. Evaluating HSPF in an arid, urbanized watershed. *Journal of American Water Resource Association* 41:477-486.

Ackerman, D. and S. Weisberg. 2003. Relationship between rainfall and beach bacterial concentrations on Santa Monica Bay beaches. *Journal of Water and Health* 1:85-89.

Ackerman, D. and S.B. Weisberg. 2006. Evaluating HSPF runoff and water quality predictions at multiple time and spatial scales. pp. 293-304 *in*: S.B. Weisberg and K. Miller (eds.), Southern California Coastal Water Research Project 2005-2006 Biennial Report. Westminster, CA.

Bicknell, B.R., J.C. Imhoff, J.L. Kittle, Jr., T.H. Jobes and A.S. Donigian, Jr. 2001. Hydrological simulation program - FORTRAN, Version 12. AQUA TERRA Consultants. Mountain View, CA.

Buffleben, M.S., K. Zayeed, D. Kimbrough, M.K. Stenstrom and I.H. Suffet. 2002. Evaluation of urban non-point source runoff of hazardous metals entering Santa Monica Bay, California. *Water Science and Technology* 45:263-8.

Daly, C. and G. Taylor. 1998. California Average Monthly or Annual Precipitation, 1961-90. <http://www.ftw.nrcs.usda.gov/prism/prism.html>1999.

Hampel, F.R. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69:383-393.

Harmel, R.D., K.W. King and R.M. Slade. 2003. Automated storm water sampling on small watersheds. *Applied Engineering in Agriculture* 19:667-674.

Izuno, F.T., R.W. Rice, R.M. Garcia, L.T. Capone and D. Downey. 1998. Time versus flow composite water sampling for regulatory purposes in the Everglades agricultural area. *Applied Engineering in Agriculture* 14(3): 257-266.

King, K.W. and R.D. Harmel. 2003. Considerations in selecting a water quality sampling strategy. *Transactions of the ASAE* 46:63-73.

King, K.W. and R.D. Harmel. 2004. Comparison of time-based sampling strategies to determine nitrogen loading in plot-scale runoff. *Transactions of the ASAE* 47:1457-1463.

King, K.W., R.D. Harmel and N.R. Fausey. 2005. Development and sensitivity of a method to select time- and flow-paced storm event sampling intervals for headwater streams. *Journal of Soil and Water Conservation* 60:323-331.

Leecaster, M., K. Schiff and L. Tiefenthaler. 2002. Assessment of efficient sampling designs for urban stormwater monitoring. *Water Research* 36:1556-1564.

Los Angeles County Department of Public Works. 2003. Data Files. Los Angeles County Department of Public Works. Alhambra, CA.

Ma, J.A., J.H. Kang, M. Kayhanian and M.K. Stenstrom. 2009. Sampling issues in urban runoff monitoring programs: Composite versus grab. *Journal of Environmental Engineering* 135:118-127.

NCDC. 2004. Hourly meteorological data. <http://lwf.ncdc.noaa.gov/oa/ncdc.html>2001.

Richards, R.P. and J. Holloway. 1987. Monte Carlo studies of sampling strategies for estimating tributary loads. *Water Resources Research* 23:1939-1948.

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

Shih, G., W. Abtew and J. Obeysekera. 1994. Accuracy of nutrient runoff load calculations using time-composite sampling. *Transactions of ASAE* 37:419-429.

Southern California Association of Governments (SCAG). 2004. 2000 Los Angeles County Land Use. SCAG. Los Angeles, CA.

Stein, E.D., L. Tiefenthaler and K. Schiff. 2005. Watershed-based sources of polycyclic aromatic hydrocarbons in urban storm water. *Environmental Toxicology and Chemistry* 25:373-385.

Stone, K.C., P.G. Hunt, J.M. Novak, M.H. Johnson and D.W. Watts. 2000. Flow-proportional, time-composited, and grab sample estimation of nitrogen export from an Eastern Coastal Plain watershed. *Transactions of ASAE* 43:281-290.

Tiefenthaler, L.L., E.D. Stein and K.C. Schiff. 2008. Watershed and land use-based sources of trace metals in urban storm water. *Environmental Toxicology and Chemistry* 27:277-287.

Tiefenthaler, L., K. Schiff and M. Leecaster. 2001. Temporal variability patterns of stormwater concentrations in urban stormwater runoff. pp. 28-44 *in*: S.B. Weisberg and D. Hallock (eds.), Southern California Coastal Water Research Project 2000 Annual Report. Westminster, CA.

ACKNOWLEDGEMENTS

We would like to thank Jay Shrake and Sean Porter of Mactec Engineering for providing cost estimates and insight in designing the microsampling approaches. We would also want to thank Fred Gonzalez of the Los Angeles Department of Public Works for information on how their sampling is conducted.