Considerations and Approaches for Monitoring the Effectiveness of Urban BMPs

Eric W. Strecker, P.E.
Woodward-Clyde
Portland, Oregon

The purposes of this paper are to 1) describe some of the problems with typical Best Management Practice (BMP) monitoring and effectiveness reporting and to 2) suggest the utilization of consistent stormwater monitoring techniques. This will allow the data collected on the effectiveness of individual best management practices (BMPs), including retrofit BMPs, to be useful for a particular site, and to also be useful for comparing studies of similar and different types of BMPs in other locations. Many BMP effectiveness studies in the past have provided only limited data useful for assessing BMP design and selection on a wide scale. This paper overviews some of the problems of past BMP effectiveness studies from the perspective of comparability between studies. It suggests some of the ways that data could be collected to make it more useful for assessing factors (such as settling characteristics of inflow solids and physical features of the BMP) that might have led to the performance levels achieved. Finally, it also discusses other considerations that affect data transferability, such as effectiveness estimations, statistical testing, etc.

Introduction

Many studies have been completed which have assessed the ability of stormwater treatment BMPs (e.g., wet ponds, grass swales, stormwater wetlands, sand filters, dry detention, etc.) to reduce pollutant concentrations and loadings. However, in attempting to summarize the information gathered from these individual BMP evaluations it is very apparent that inconsistent study methods and reporting make wider scale assessments difficult. For example, individual studies often included the analysis of different constituents and utilized different methods for data collection and analysis. These differences alone contribute significantly to the range of BMP effectiveness reported. This makes assessing what other factors may have contributed to the variation in performance almost impossible.

In one review of the use of wetlands for stormwater pollution control (Strecker et al., 1992), a summary of the literature on performance of wetland systems and the factors that may have led to the reported pollutant removals was prepared. The literature was inconsistent with respect to the constituents analyzed and the methods used to gather and analyze data. A number of pieces of information, if collected and recorded, would have improved the ability to evaluate the effectiveness of stormwater wetlands as BMPs and facilitated the transfer of that knowledge into better design practices. Urbonas (1994 and 1995) and Strecker (1994) summarized the information that should be recorded about the physical, climatic, and geological parameters which likely affect the performance of a BMP, and considerations regarding sampling and analysis methods. This paper presents 1) a suggested list of constituents for analysis along with recommendations for reporting data, 2) methods of reporting pollutant removal efficiencies, 3) a brief discussion of statistical approaches to selecting the number of samples needed, 4) methods for including detection limit data, 5) sample collection considerations, and 6) the need for dry weather assessments.

BMP Performance Study Inconsistencies

Studies of BMP effectiveness have utilized significantly different:

- Sample collection techniques (e.g., from sample collection types (grab, composite, etc.), flow measurement techniques, to how the sample was composited, etc.);
- Constituents, including: chemical species, methods (detection limits), form (e.g., dissolved vs. total, vs. total recoverable, etc.), and treatment potential;
- Data reporting on tributary watershed and BMP design characteristics (e.g., tributary area or watershed attributes such as percent impervious, land use categories, rainfall statistics, etc.);
- Effectiveness estimation (at least four techniques have been utilized to assess effectiveness which can cause significant differences in pollutant removal reporting, with the same set of data), and potential alternatives to reporting just concentration/loading reductions; and
- Statistical validation of results (typical lack of statistical tests to determine if the reported removal efficiency

can in fact be shown to be statistically different than zero).

Any of the above topics would require an in-depth discussion beyond the scope of this paper to fully explain. Therefore, this paper will present a brief overview of each of these and some potential solutions to improving how data is collected. EPA together with ASCE is currently developing a set of protocols and a database on BMP performance studies with the purpose of improving the consistency of BMP monitoring information. This project includes:

- Developing Protocols for BMP Monitoring
- Conducting an evaluation of existing information to assist the EPA Wet Weather FACA and contribute to EPA's Stormwater Toolbox (as identified in Draft Phase II Stormwater Regulation Preamble)
- Developing a data base on BMP performance studies

The overall goal is to improve the BMP effectiveness information base to:

- Develop information to improve designs
- Improve performance information

The data base specifies a chosen set of reporting information, but does not tell how to develop such information. For example, it does not specify what a flow-weighted composite sample is and how it should be collected. The next step beyond the EPA protocols and data base effort should be a guidance document on monitoring data collection strategies and techniques to improve their consistency and transferability. It should be recognized that with the development of the database and the protocols, it will be a number of years (5 to 10) before significant new studies on BMPs are conducted utilizing the protocols to allow for a more rigorous evaluation of BMP selection and design factors.

Sample Collection Techniques

The differences among sample collection techniques alone is enough to make comparing different studies questionable. These include differences among how flows are measured to how samples are composited to formulate an "event mean concentration." Some studies have utilized grab samples, and the results of these studies in evaluating BMP performance are limited. Typically studies will include the collection of flow-weighted composite samples (either automated or hand collected). These studies involve various techniques (often not reported very well) for measuring flows. The flow measurements themselves are subject to a large variation.

The Federal Highway Administration is currently conducting a study of monitoring techniques for characterizing stormwater runoff hydrology and water quality from highways. The study, being completed by Woodward-Clyde, included a component conducted by the USGS (Waschbusch and Owens, 1998) which addressed the potential differences in flow measurement techniques in a pipe system in Madison, WI. An in-depth dye-dilution method was utilized to calibrate a Palmer-Bowlus flume with a bubbler pressure measurement. The study evaluated 23 flow measurement techniques including commercially available packages and individual component systems.

Figure 1 is a summary of the results of flow measurements, showing the average percent differences from the calibrated flume. These data summarize 50 storm events which were measured over a 6-month period. As the figure demonstrates, the error in flow measurements is easily on the order of plus or minus 25% over a range of storms. The flow measurements for individual storms varied even more. If samples are composited based upon flows (either using automated or using grab samples), they are subject to an error in collection times (for automated systems) or in composited amounts (grab sample composited) and therefore could result in errors in estimates of event mean concentrations (especially for constituents which vary over the course of a storm event). It should be strongly noted that these results are for one site only and should not be interpreted as indicative of how any particular system identified might perform at another site. It is imperative that researchers thoroughly evaluate potential flow measurement alternatives and implement the method that will result in the best information possible.

Another aspect of the study addressed how many samples should be collected to compile a "flow-weighted" composite sample. Figure 2 demonstrates the large variability in sampler bottle configurations. These configurations often drive researchers into selecting the number of "grab" composite samples to collect. For example, in the NPDES monitoring for Texas (Brush, et al., 1994), the chosen strategy was to collect one sample into each bottle of the 8-bottle configuration (this was successful if it rained sufficiently). In the Portland and Eugene NPDES Sampling (WCC 1993a and WCC 1993b), an attempt was made to collect 24 "grab" samples during the course of an event. Figure 3 shows a typical storm event from the Portland program and specifically the points at which a sample was collected. From the variability in flows observed, one can surmise the pollutant concentrations were also fluctuating extensively (later confirmed by within-storm sampling). Having only eight samples during this event may not have accurately characterized the event mean concentration (EMC). Collecting three times the samples to "construct" a flow-weighted sample would appear to reduce the chances of anomalies (variability) during a storm event influencing the overall estimate of the average concentration. Early results from our FHWA study indicate that one should attempt to collect at least 12 to 16 individual samples to form a composite sample.

The study also has evaluated the potential effects of sample lift (e.g. pumping up from underground or from stream bottoms) and has found that the newer samplers

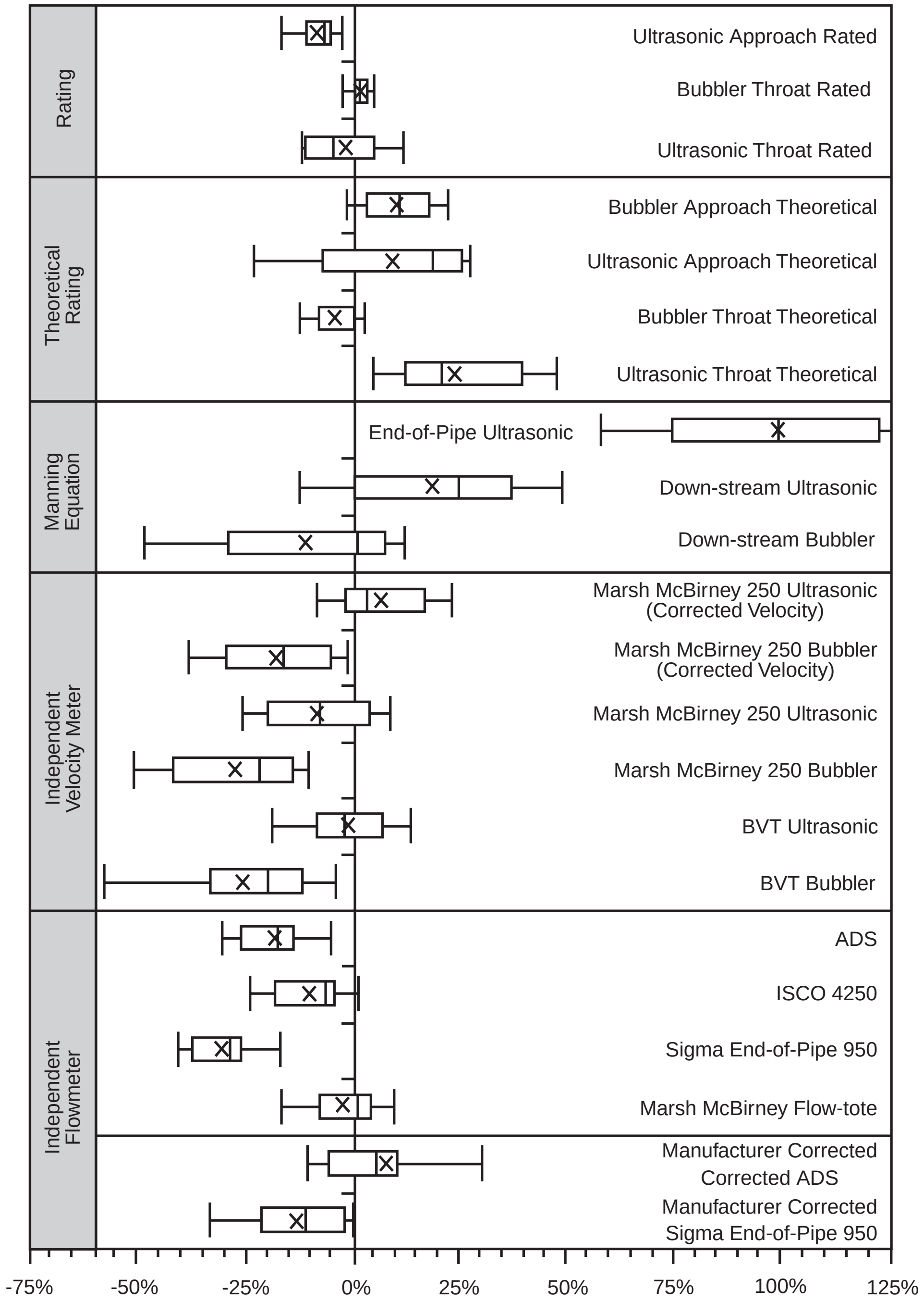


Figure 1. Boxplot of the percent differences between total storm volumes computed using various flow estimation methods and the total storm volume of the bubbler approach rated discharge (bold line at 0%). (Waschbusch and Owens, 1998).

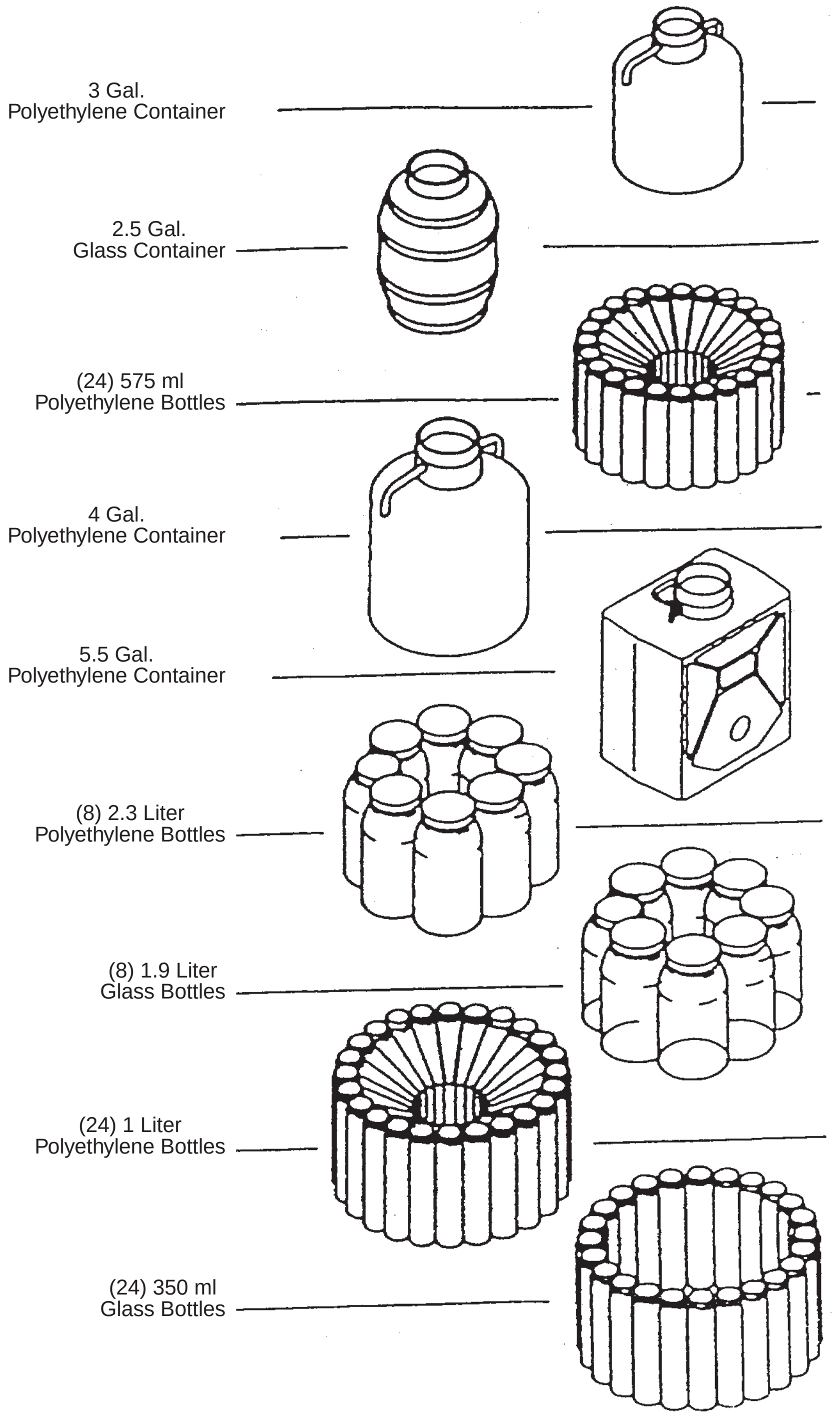


Figure 2. Typical automated sampling bottle configuration options.

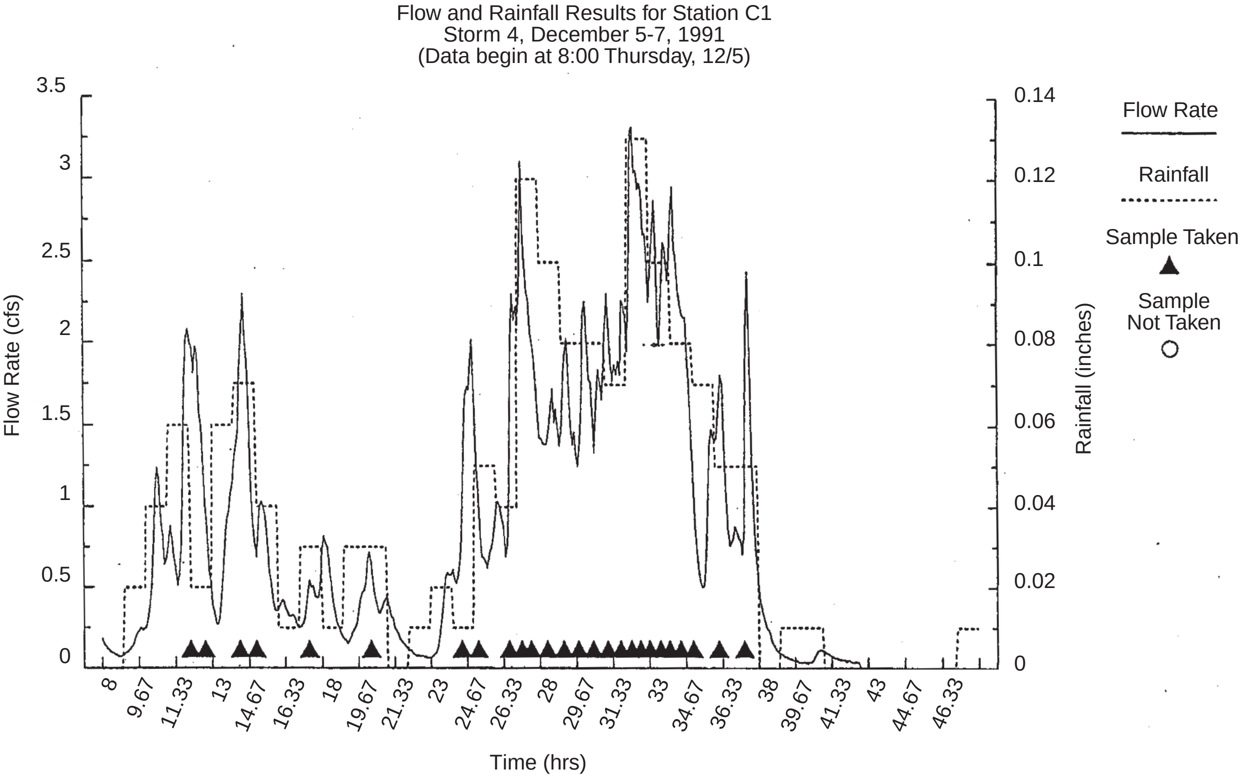


Figure 3. Typical hydrograph indicating measured rainfall, runoff, and water sample collection times from automated flow and water quality sampling for the Portland NPDES Stormwater Monitoring Program.

(with stronger pumps) do not appear to cause any separation of suspended solids as they are lifted up to 20 feet. At the end of the study, a guidance document on sampling of highway runoff will be developed. These are just some of the numerous differences in sampling methods that could lead to differences in results between BMP studies.

Constituents Assessed

A very wide variety of pollutants have been analyzed in both BMP studies and characterizations studies. The EPA protocols study has developed a recommended set of constituents for BMP testing programs. These were developed from the review of previous studies and an understanding of costs and likelihood of providing meaningful results. Below is a discussion of how these constituents were selected (adapted from Strecker, 1994).

Since NURP and prior to the Phase I Stormwater NPDES monitoring programs, there have been a number of studies which continued to assess pollutant concentrations in stormwater runoff. These included the Federal Highway Administration's highway runoff program (Driscoll et al., 1990) and some selected studies done in a few locations. These studies typically were not consistent with the standard NURP protocols. Based upon the 1987 amendments to the Clean Water Act, EPA required operators of municipal separate storm drainage systems that served populations of over 100,000 to collect flow-weighted composites at a minimum of five stations to characterize residential, commercial, and industrial runoff quality. Only a few additional parameters have been identified as "problems" in stormwater, based upon these post-NURP studies (this despite the improved analytical methods that have become available for conducting laboratory analyses). In addition, NURP focused primarily on residential and commercial land uses, while NPDES testing included industrial land uses which were suspected of having more pollutants present.

However, there has not been a comprehensive review by EPA or others of the newly collected stormwater information to assess the results of requiring the analysis of over 130 constituents, including priority pollutants. This type of review is needed. EPA's requirements included monitoring three storms at selected stations. This number of storms is only useful for identifying potential problem pollutants. Statistically, these are not enough data to perform a meaningful regional or other factor analyses of urban stormwater concentrations, although they could provide useful information on rates of detection. This analysis would be helpful in selecting constituents for BMP monitoring.

The choice of constituents to include as "standard pollutants" is a subjective one. As an example, some would argue that cost should be a primary consideration; others would say that it should not. In making the recommended list of monitoring constituents, the following characteristics were considered:

- The pollutant is prevalent in typical urban stormwater at concentrations that could cause water quality impairment.
- The analytical test can be related back to potential water quality impairment.
- Sampling methods for the pollutant are straightforward and reliable for a moderately careful investigator.
- Analysis of the pollutant is economical on a widespread basis.
- The pollutant is one for which treatment is a viable option.

Not all of the pollutants recommended fully meet all of the factors listed above; however, the factors were considered in the recommendations. When developing a list of pollutant analyses for an individual BMP evaluation, it is important to consider the upstream land use activities. The parameters recommended below are present and of concern in "typical" urban stormwater.

The Nationwide Urban Runoff Program (NURP) (EPA, 1983), which included monitoring of land use runoff and BMP performance at over 28 cities nationwide, adopted consistent data collection methods and analytical parameters. Results from the NURP program could be used to evaluate similarities and differences in pollutant concentrations in urban stormwater from different and similar land uses, and could be used to explain what might be causing these differences. The following pollutants were adopted by NURP as "standard pollutants characterizing urban runoff":

TSS	Total suspended solids
BOD	Biochemical oxygen demand
COD	Chemical oxygen demand
TP	Total Phosphorus
SP	Soluble phosphorus
TKN	Total Kjeldahl nitrogen (as N)
$NO_2 + NO_3$	Nitrate + nitrite (as N)
CU	Copper
PB	Lead
ZN	Zinc

Oil and grease was not included because of the difficulty in obtaining representative samples. On a less consistent basis, NURP also monitored for pollutants including other metals, dissolved metals, semi-volatile organics, volatile organics, pesticides, and herbicides.

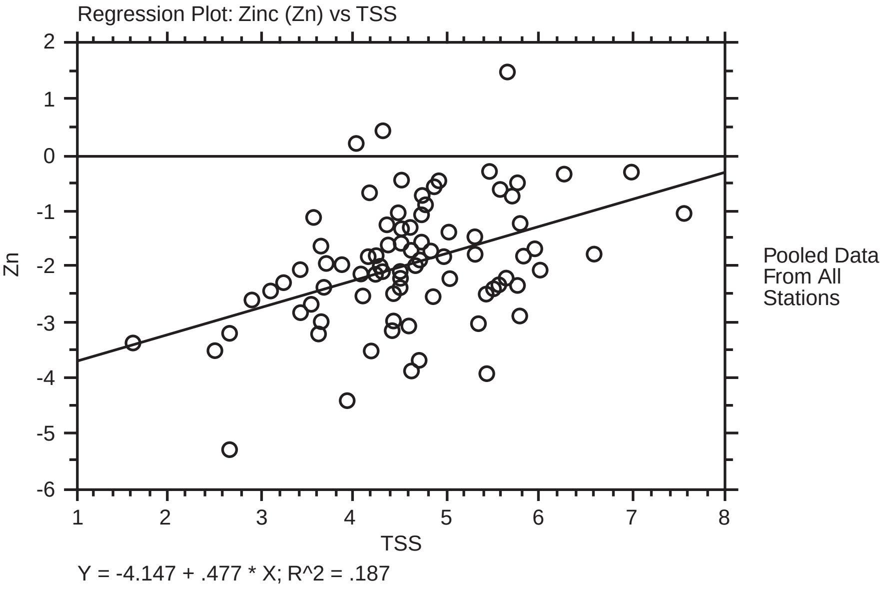
Presented below is a brief discussion, by group, of the pollutants that are recommended to be included in a base list, then several that may occasionally be recommended.

<u>Total Suspended Solids (TSS)</u>. The term "suspended solids" is descriptive of the organic and inorganic particulate matter which is of a size and type that allows the particles to stay suspended in water. The solids load in a waterbody is influenced by a number of factors including but not limited to: particle sizes, stream flows, climate, geology, and vegetation of each drainage system. The conditions under which suspended solids are considered a pollutant is a matter of definition. In general, suspended solids are considered a pollutant when they significantly exceed natural concentrations and have a detrimental effect on water quality and/or beneficial uses of the water body.

Suspended sediments are often used as a surrogate for other contaminants which bind or adsorb easily with fine particulate matter, including heavy metals. Although TSS is often highly correlated with other parameters, it is generally not a strong enough correlation to eliminate the need to address other parameters specifically. Figure 4 shows the relationship between TSS and zinc for pooled stormwater runoff monitoring data from all ten stations monitored in Portland, Oregon for the NPDES program (WCC, 1993a) and from the seven stations that were from piped systems. Although the relationship is statistically significant (R^2 of .38 for piped stations), it does not explain a significant amount of the variability. Similar results were found for almost all other parameters. It should be noted that for individual stations, the relationships between TSS and many pollutants were sometimes much higher, but this would mean that one would have to monitor enough times to establish the relationship. Therefore, TSS does not appear to be a good predictor of other pollutants, without significant data collected from each station. However, TSS is one good indicator of pollutant removal efficiency (e.g., because of the tendency for many pollutants to be associated with fine particulates) and should be included in any evaluation of BMP performance.

Many BMPs rely on sedimentation as the primary pollutant removal mechanism. It is recommended that samples also be analyzed for some measure of the expected settling rate (treatment potential) of TSS. The performance of a BMP that relies on sedimentation and even filtering can be greatly affected by the particle sizes and densities present in the influent. If the influent TSS is characterized by very small particle sizes, and therefore slow settling velocities, it will be much more difficult to treat. The settleability of influent solids has not been adequately addressed in performance comparisons, and may be one of the significant reasons that measured performance varies so highly from similar BMP to BMP.

For consideration, the particle size distribution in street dirt found in Sartor and Boyd (1972), as shown in Table 1,



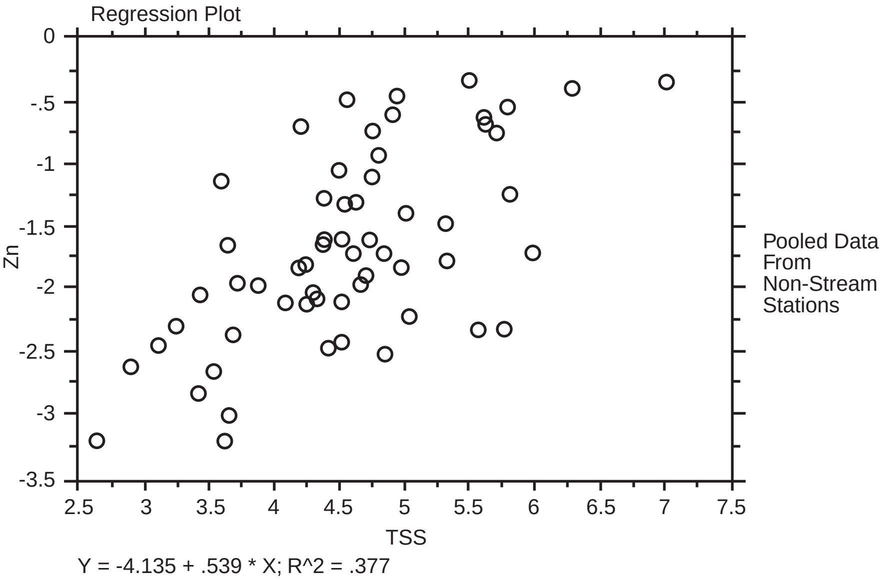


Figure 4. Natural logarithm regression plots of zinc (Zn as ln(mg/l)) vs. total suspended solids (TSS) for pooled Portland, OR stormwater monitoring event mean concentration data. (Woodward-Clyde Consultants, 1993a)

might be an appropriate gauge of the "treatment potential" of stormwater. As Table 1 indicates, these distributions vary considerably from city to city and likely from site to site. One can easily surmise that if testing were performed on similar catchments and BMP designs, that there could be a large difference in BMP performance results from these sites just due to the particle size differences alone. Another potential measure of the treatment potential would be information from settling column tests as those discussed by EPA in its manual on combined sewer overflow control (EPA, 1993a).

Oxygen Demand. Oxygen demand refers to the amount of oxygen that will be consumed by biological or chemical reactions involving organic compounds. The decomposition of biodegradable materials by natural soil and water bacteria draws upon the dissolved oxygen resources of a water body. This process is countered by natural re-aeration processes that occur in all water bodies to varying degrees. Significant reductions in dissolved oxygen concentrations can result when the demand rate exceeds the rate of replenishment through re-aeration. In general, moderately high dissolved oxygen content is necessary for the maintenance of healthy aquatic ecosystems. The relationship of oxygen-consuming discharges to the amount of dissolved oxygen in a receiving water body, therefore, is fundamental to the maintenance of environmental quality in natural water bodies. However, the tests available for

Table 1. Particle Size Fractions of Street Dirt from Selected Locations.

Size Ranges	Milwaukee	Bucyrus	Baltimore	Atlanta	Tulsa
>4800 μ	12.0%	–%	17.4%	–%	–%
2000 - 4800 μ	12.1	10.1	4.6	14.8	37.1
840 - 2000 μ	40.8	7.3	6.0	6.6	9.4
246 - 840 μ	20.4	20.9	22.3	30.9	16.7
104 - 246 μ	5.5	15.5	20.3	29.5	17.1
43 - 104 μ	1.3	20.3	11.5	10.1	12.0
30 - 43 μ	4.2	13.3	10.1	5.1	3.7
14 - 30 μ	2.0	7.9	4.4	1.8	3.0
4 - 14 μ	1.2	4.7	2.6	0.9	0.9
>4 μ	0.5	–	0.9	0.3	0.1
Sand %, 43 - 3800 μ	92.1	74.1	82.1	91.9	92.3
Silt %, 4 - 43 μ	7.4	25.9	17.1	7.8	7.6
Clay %, <4 μ	0.5	–	0.9	0.3	0.1

Note: μ = microns
Source: Sartor and Boyd, 1972

assessing oxygen demand are not straightforward indicators of potential problems.

Biochemical Oxygen Demand (BOD). The 5-day BOD test provides an indirect measure of the quantity of biologically degradable organic matter in water in terms of the amount of oxygen required by microorganisms to oxidize it to carbon dioxide and water. The BOD test is quite variable. A number of factors can affect results, including the quality of the seed culture utilized in the test. The BOD test can also be inhibited by toxicants in the sample, which may react differently once the runoff mixes with the receiving water. The levels of BOD that are normally found in urban stormwater are near detection limits for the BOD test. Therefore, they are subject to wide variation. Therefore BOD has not been recommended as a parameter. Instead, TOC (Total Organic Carbon) has been identified as a more consistent measure of available organic material, which could be contributing to oxygen demand.

Chemical Oxygen Demand (COD). The COD test provides a more rapid and consistent measure of oxygen demand than BOD tests. The consumption of oxygen from an introduced strongly oxidizing chemical agent is measured by this test. As a result, it typically measures appreciably higher levels of oxygen demand than will be produced by biological decomposition because it oxidizes some organic compounds that are not biodegradable, and may also react with inorganic compounds as well. In urban stormwater, for example, COD levels are typically found to be about 8 to 10 times greater than BOD levels. COD measures a “maximum possible,” but not probable, oxygen demand.

Nutrients. Nutrients are necessary for the growth and support of biota in natural water systems. Excessive quantities can result in the over-stimulation of biological growth and the creation of objectionable water quality conditions (eutrophication). Some forms of nutrients can also be toxic (e.g., ammonia). In general, the most important nutrient factors causing an acceleration in algal production are nitrogen compounds and phosphorus.

Nitrogen. Nonpoint sources of nitrogen include lawn fertilizers, leachate from waste disposal in dumps or sanitary landfills, atmospheric fallout, nitrite discharges from automobile exhausts and other combustion processes, natural sources such as mineralization of soil organic matter, and farm-site fertilizers and animal wastes. Many water treatment methods have no significant effect on nitrate removal from water (Dunne and Leopold, 1978).

Three forms of nitrogen have been analyzed extensively in stormwater runoff water quality studies. These are nitrite plus nitrate ($NO_2 + NO_3$), ammonia nitrogen (NH_3), and total Kjeldahl nitrogen (TKN). The latter, named after the analytical test procedure, provides a measure of ammonia and organic nitrogen forms that are present. The first ($NO_2 + NO_3$) provides a measure of the inorganic nitrogen. There is usually very little nitrite in stormwater. Nitrate (NO_3) is very mobile and is usually difficult to treat utilizing stormwater BMPs. Ammonia nitrogen can be toxic to aquatic life. It can be assessed for toxicity to aquatic life with data on pH and temperature. The inorganic (NO_2 + NO_3) and ammonia nitrogen are recommended. All forms should be reported as mass of nitrogen (N).

Phosphorus. Phosphorus is used by algae and higher aquatic plants and may be stored in excess of use within plant cells. With decomposition of plant cells, some phosphorus may be released immediately through bacterial action for recycling within the biotic community, while the remainder may be deposited with sediments.

Phosphorus enters waterways from many of the same sources as nitrogen. Domestic sewage contains significant concentrations of phosphorus which are contributed by detergents and human wastes. Primary and secondary treatment processes normally remove only about 20 to 30% of this element from sewage (Dunne and Leopold, 1978). Fertilizers and the erosion of soils rich in phosphorus can also be a potential source.

Three forms of phosphorus have been somewhat routinely analyzed in stormwater runoff studies. These include total phosphorus (TP), soluble phosphorus (SP), and ortho-phosphate (OP). Ortho-phosphate indicates the phosphorus that is most immediately biologically available. Soluble phosphorus includes both the ortho-phosphate and a fraction of the organic phosphorus. Most all of the SP is usually OP, however. Total phosphorus includes phosphorus in the forms that may not be as readily biologically available plus the forms discussed above. TP and OP are recommended for inclusion in a monitoring program, as they characterize both the total and bioavailable forms of phosphorus. All forms should be reported as mass of phosphorus (P).

Metals. Heavy metals such as copper, lead, and zinc are naturally released in very small quantities by the weathering of exposed soils and mineral deposits, corroding metal surfaces, decomposing paints, and certain corrosion-control compounds. Heavy metals tend to have comparatively low solubilities and are often mobilized by forming soluble complexes with humic materials or by becoming attached to clay particles. Heavy metals have been consistently identified as the most significant toxics found in urban stormwater and often exceed water quality criteria for aquatic life.

These metals are present in the biosphere as trace elements and are micronutrients necessary for plant and animal growth. Heavy metals are of concern because elevated concentration levels of soluble forms in natural water bodies can produce toxic effects in biota. Sources include domestic and industrial point-source discharges, urban stormwater runoff, and direct atmospheric deposition. In this paper, copper (Cu), lead (Pb), zinc (Zn), and cadmium (Cd) have been recommended for inclusion in a monitoring program because stormwater runoff water quality studies conducted at many urban locations have indicated that these metals are almost always present, and are at concentrations which tend to be elevated, relative to other heavy metals. They also can be used as surrogates for other heavy metals, as they tend to display the range of transport characteristics for heavy metals. However, other heavy metals should be analyzed if there are known sources of significant quantities of these metals in influent flows.

It is recommended that both the total and dissolved form of each be analyzed. Based upon EPA's recommendation, the dissolved fraction should be compared to water quality criteria, with modifications to the criteria as noted in EPA (1993b). To compare data to criteria, hardness should be measured for each sample. Too often, metals data are compared to criteria using an average hardness value not directly associated with the monitoring, and not associated with storm events. In the Williamatte Valley of Oregon, stormwater sampling has shown that hardness values during storm events are quite low, which results in low criteria values.

Total concentrations are valuable in assessing the overall reduction of the heavy metal in both soluble and particulate forms. There is a concern about the long-term bioavailability of these metals in sediments and sediment standards are beginning to be developed and implemented.

When conducting these tests, it is recommended that low detection limits be achieved. For copper, lead, and zinc, the detection limit should be 1 μg/l and for cadmium 0.2 μg/l. This will minimize problems with analyses that include below detection limit data, which can severely impact performance evaluations. Special "clean" procedures will be necessary to achieve low detection limits, both in the laboratory and in the field.

Too often, BMP effectiveness for metals is estimated based upon data that is very near or below detection. This is troublesome when both the inflow and outflow concentrations are at or near detection, and effectiveness is based upon a storm-by-storm comparison of loads or concentrations. It is recommended that if both the influent and effluent concentration are within five times the method detection limit, the pollutant data pair not be considered in the effectiveness analysis if a storm-by-storm method is used. If statistical characterizations of the inflow and the outflow concentrations are utilized to assess effectiveness and some of the data are below detection, appropriate techniques should be utilized. Driscoll et al. (1990) describes a method to address detection limit data. The setting of below-detection values to 0 or 1/2 the detection limit or the detection limit, will typically lead to an underestimation of the mean.

Oil and Grease. Oil and grease is a prevalent constituent in urban runoff and often exceeds discharge limits set by states (such as 10 mg/l in Oregon for industrial stormwater permits). In a study of oil and grease concentrations in urban runoff in Richmond, California, Stenstrom et al. (1984) found that oil and grease concentrations in runoff from commercial properties and parking lots are about three times higher than from residential and open areas. The NURP program did not address oil and grease as a standard constituent. Accurately measuring oil and grease is very difficult due to its affinity for coating sampling bottles and sampling tubes and its highly non-uniform distribution in the water column (except in the most turbulent situations). Other tests include total petroleum hydrocarbons, which measure the petroleum based fraction of oil and grease. Other sources of oil and grease include animal and vegetable. For BMPs which are designed to address oil and grease, it is suggested that some multiple, within a storm, grab sample analyses would be appropriate. For most BMPs, it is recommended that the parameter be optional. If completed, the TPH evaluation is recommended as the most appropriate measure to gauge effectiveness of a BMP at reducing man-induced sources of petroleum oil and greases.

Pesticides/Herbicides. Pesticides and herbicides are regularly detected in urban runoff. However, the number of constituents usually detected is low and most often at levels below available criteria. In Portland, Oregon (WCC, 1993a) the frequency of detection of pesticides herbicides was less than 1% of all the pesticides and herbicides tested. However, the city has noted locations where pesticide concentrations in sediments are high. This could indicate that the problem might be due to misuse or dumping, rather than a general stormwater problem. Although it is possible that pesticides accumulate in sediments from low concentrations in stormwater, some regional assessments of the effectiveness of source control measures (education, identification and elimination of dumping problems) are needed. The Alameda County, CA monitoring program (Cooke and Lee, 1993) and other studies have recently identified that the pesticide Diazinon may be a primary cause of toxicity at very low concentrations (below 8140 method detection limits) to *cerodaphrin dubia* in receiving streams in the south bay area of San Francisco. More research is needed to further define the level of this problem in relation to the actual instream biota, rather than test organisms. At this time, I would not recommend including the pesticide in a standard list, but research studies on the magnitude of the problem and the effectiveness of BMPs on these pesticides should be performed. Due to the low values at which these constituents can cause problems, it would be very difficult to assess BMP performance on a wide-scale basis. For example, it may be more appropriate to eliminate or control the use of Diazinon rather than research BMP effectiveness on concentrations that are below 1 ppb.

Volatile and Semi-Volatile Organics. These pollutants have not generally been detected at a high frequency and in quantities that exceed available criteria [with the exception of Polynuclear Aromatic Hydrocarbons (PAHs), which are discussed separately]. In the recent City of Portland and Eugene sampling programs (WCC, 1993a and 1993b) detection rates were less than 2% of all the tested constituents and below all available criteria. These parameters are not recommended for general analysis unless a BMP effectiveness study is being conducted in an industrial area suspected or known to have elevated levels of organics.

Polynuclear Aromatic Hydrocarbons. The carcinogenic properties of PAHs have generated increased interest in the study of their sources, transport, fate, and aquatic toxicity. Major sources include the combustion of fossil fuels, uncombusted petroleum products (fuels, etc.), and natural and man-caused fires. PAHs have recently been analyzed utilizing detection levels that are significantly below those achieved utilizing the standard semi-volatile organic scans (WCC, 1993a and 1993b; Cooke and Lee, 1993). These tests (GC-MS methods at the nanogram per liter level) have shown that PAHs in stormwater are above human consumption criteria by significant amounts (up to over 100 times). However, these tests are specialized (only a few laboratories provide this level of analysis) and expensive (about $500 to $600 per analysis). In addition, there are no criteria for aquatic life, and toxicity identification evaluations performed in the San Francisco Bay Area have not identified PAHs as the source of toxicity in either developed land-use runoff or in stream stations. For these reasons, PAHs are not recommended for the standard list of constituents to be monitored. However, because of their carcinogenic nature and their tendency to bioaccumulate, new studies may identify potential long-term aquatic life impacts that may require reevaluation of this recommendation.

Data Reporting

Practical and technical data reporting considerations, including consistent formatting of data, the clear indication of QA/QC results, standard comparisons to water quality criteria, reporting of tributary watershed characteristics, and BMP design information would facilitate data usefulness. The last two items are considered critical for evaluation of what contributed to BMP effectiveness in one location over another.

Data Formatting. It is recommended that all constituent concentration data be reported as event mean concentrations (EMCs). Table 2 is an example format for reporting storm event EMCs. It indicates the date of the storm, the EMC value for each sampling point, the data that are estimates based upon QA/QC evaluations, method used for analysis, and detection limit achieved. Also included are summary statistics of the EMCs. These statistics should be based on use of the lognormal distribution. The NURP and FHWA studies (EPA, 1983; Driscoll et al., 1983) identified the lognormal distribution as suitable for characterizing EMC distributions. An example of the variability in data is shown in Figure 5. The figure shows a log-probability plot for total copper collected at a commercial land use station. The event mean concentrations ranged from 6 to 70 µg/l. This high degree of variability is why proper statistical techniques should be employed to evaluate whether a measured difference between BMP before/after or input/output is truly different.

The inclusion of outlet data as a part of any paper or report will allow comparisons of typical outlet concentrations and may allow the determination of the lowest or average expected concentration from a particular type of BMP. For example, it may be that wet ponds may only be able to treat to some minimum concentration range at the outlet and the “effectiveness” is greatly impacted by the inlet concentrations.

Quality Assurance/Quality Control (QA/QC). All monitoring studies should include a QA/QC program. The results of the QA/QC program should be reported in monitoring study reports and summarized in papers. It is especially important to discuss when data are characterized as estimates due to QA/QC results and when detection limits were affected. Too often this information is not included.

Table 2. Example Data Reporting Table from Eugene NPDES Monitoring Summary Report (Woodward-Clyde Consultants, 1993b)

Chromium	(mg/L)			Method EPA 7191				Receiving Water Quality Criteria**
Storm Event	Date Sample	R-1	C-1	I-1	I-2	M-1	M-2	Detection Limit
#1	9/23/92	0.034	0.016	0.031	0.008		0.003	*0.001*
#2	12/5/92	0.005		0.005	0.003	0.001		*0.001*
#3	12/16/92	0.004	0.004	0.008	0.009		0.003	*0.001*
#4	1/19/93	0.004	0.012	0.019	0.011	0.008	0.004	*0.001*
#5	3/14/93	0.003	0.006	0.020	0.017	0.004	0.007	*0.001*
#6								
#7								
#8								
	Median	0.006	0.008	0.014	0.008	0.003	0.004	
	COV	1.27	0.70	0.86	0.71	–	0.42	
	Mean	0.010	0.010	0.018	0.010	–	0.004	

Copper	(mg/L)			Method EPA 6010				Receiving Water Quality Criteria**
Storm Event	Date Sample	R-1	C-1	I-1	I-2	M-1	M-2	Detection Limit
#1	9/23/92	0.081	0.130	0.071	0.016		0.019	0.001
#2	12/5/92	0.004		0.01	0.01	0.009		0.001
#3	12/16/92	0.011	0.016	0.037	0.03		0.009	0.004
#4	1/19/93	0.009	0.046	0.076	0.034	0.027	0.012	0.003
#5	3/14/93	<0.030	0.027	0.034	0.025	0.020	<0.030	0.004
#6								
#7								
#8								
	Median	0.012	0.040	0.037	0.021	0.017	0.013	
	COV	2.03	1.11	0.97	0.54	–	–	
	Mean	0.030	0.060	0.051	0.024	–	–	

Results expressed as mg/L (ppm) unless otherwise noted. COV is the Coefficient of Variation. ** Criteria are hardness dependent.
“nd” means none detected at or above the detection limit listed. If no value is shown, the lab analysis was not performed.
Summary statistics are based on the assumption that the samples of EMCs are lognormally distributed.
Italicized values are considered estimates due to QA/QC review but are included in the calculations.

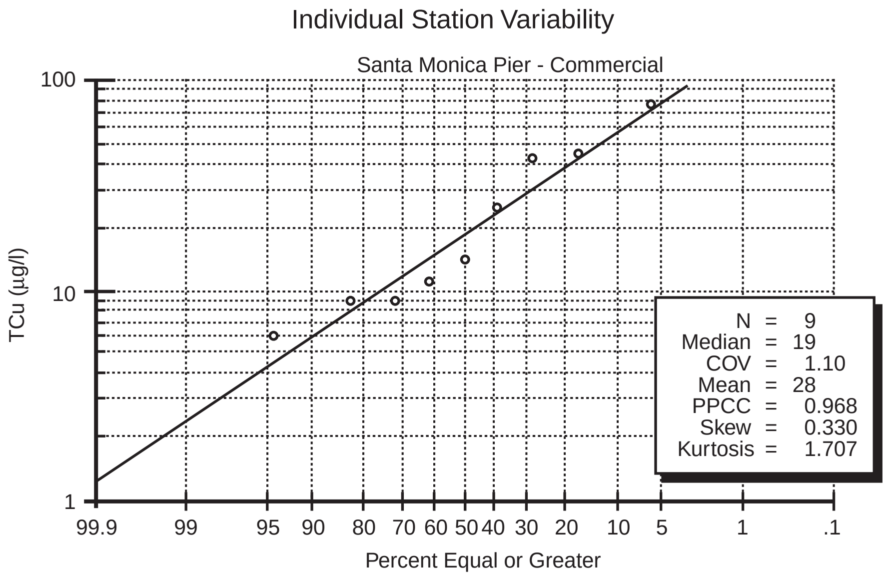


Figure 5. Example log probability plot of storm event mean concentrations from data collected by the Los Angeles County Stormwater Monitoring Program at the Santa Monica Pier (Santa Monica, CA) Commercial Land Use Station.

Comparisons to Water Quality Criteria. Another method to gauge effectiveness could be to monitor how the BMP effects the number of times that criteria are exceeded in both the inflow and the outflow, to assess how the BMP reduces (or does not reduce) the frequency of storm events where water quality criteria are exceeded. For heavy metals analyses, it is recommended that hardness be collected for all storms monitored and that comparisons to criteria be made utilizing the dissolved fraction with the computed aquatic criteria as modified by EPA (1993b). Figure 6 presents an example presentation of metals exceedances for data collected in Portland, OR (WCC 1993a). These data could be compared to BMP data for exceedances to determine whether or not a BMP was actually reducing potential toxicity.

Watershed BMP Design Parameters. Urbonas (1995) described information that should be collected regarding the physical, climatic, and geologic parameters, which include watershed and BMP design characteristics that could likely affect the performance of a BMP. Table 3 (Strecker and Urbonas, 1995) presents a summary of these parameters. More detailed and updated lists will be published upon completion of the EPA study referenced earlier.

Estimation of Pollutant Removal Effectiveness

BMP pollutant removal effectiveness estimations are not straightforward and a wide variety of methods have been employed. Martin and Smoot (1986) discussed the following three types of methods to compute efficiencies:

- The first method employs an efficiency ratio (ER), which is defined in terms of the average event mean concentration (EMC) of pollutants from inflows and outflows, thus:

$$ER = 1 - \frac{\text{Average outlet EMC}}{\text{Average inlet EMC}}$$

- The second method is based on the summation of loads (SOL) of pollutants removed during the monitored storms, thus:

$$SOL = 1 - \frac{\text{Sum of outlet loads}}{\text{Sum of inlet loads}}$$

- The third method of determining efficiency, developed by Martin and Smoot (1986), defines the ratio as the slope of a simple linear regression of inlet loads and outlet loads of pollutants. The equation for the regression of loads (ROL) efficiency is thus:

$$\text{Loads in} = ß \cdot \text{Loads out}$$

where ß equals the slope of the regression line, with the intercept constrained at zero.

The ER and SOL methods assume that monitored storms include samples representative of all storms that occur. The SOL method assumes that enough samples were collected so that any significant input loads or output loads were not missed. They are different in that one gauges effectiveness in terms of concentration reduction, while the other gauges effectiveness in terms of load of pollutant removed. The ROL method assumes that the treatment efficiency is the same for all storms, which is likely not the case.

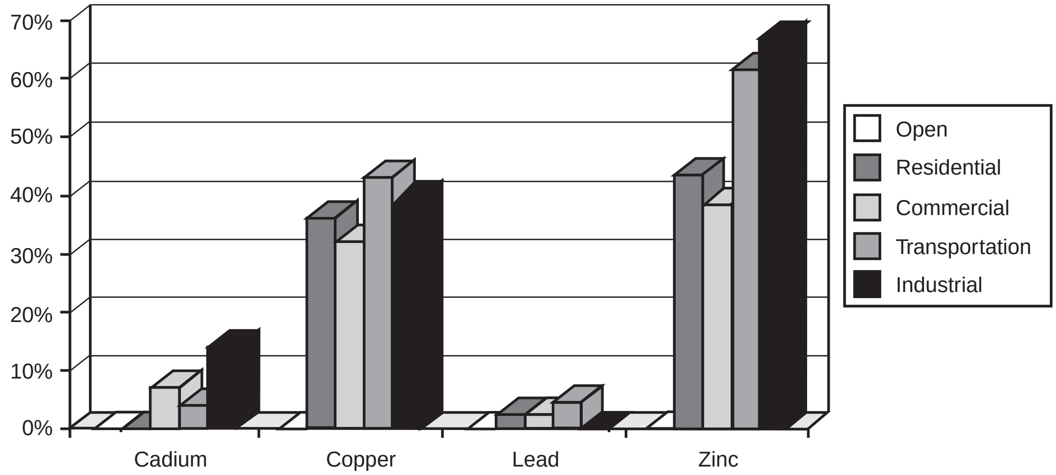


Figure 6. Frequency of water quality criteria exceedances of Oregon urban stormwater data collected for the Municipal Stormwater NPDES Programs.

Table 3. Parameters to Report with Water Quality Data for Various BMPs

Parameter Type	Parameter (1)	Retention (Wet) Pond (2)	Extended Detention Basin (3)	Wetland Pond Basin (4)	Grass/Swale Wetland Channel (5)	Sand/Leaf Compost Filter (6)	Oil & Sand Trap (Vault) (7)	Infiltration and Percolation (8)
Tributary Watershed	Tributary watershed area	•	•	•	•	•	•	•
	Total tributary watershed impervious percentage	•	•	•	•	•	•	•
	Percent of impervious area hyd. connected	•	•	•	•	•	•	•
	Gutter, sewer, swale, ditches in watershed?	•	•	•	•	•	•	•
	Land use types (res, comm, ind. open) and acreages	•	•	•	•	•	•	•
General Hydrology	Average storm runoff volume	•	•	•	•	•	•	•
	50th percentile storm runoff volume	•	•	•	•	•	•	•
	Coefficient of variation of runoff volumes	•	•	•	•	•	•	•
	Average daily base flow volume	•	•	•	•	•	•	•
	Average runoff interevent time	•	•	•	•	•	•	•
	50th percentile interevent time	•	•	•	•	•	•	•
	Coefficient of variation of interevent times	•	•	•	•	•	•	•
	Average storm duration	•	•	•	•	•	•	•
	50th percentile storm duration	•	•	•	•	•	•	•
	Coefficient of variation of storm durations	•	•	•	•	•	•	•
	2-year flood peak velocity				•		•	
	Depth high groundwater of impermeable layer		•	•				•
Water	Water temperature	•	•	•	•	•	•	•
	Alkalinity, hardness and pH	•	•	•	•	•	•	•
	Sediment setting velocity distribution, when available	•	•	•	•	•	•	•
	Facility on- or off-line?	•	•	•	•	•	•	•
	If off-line, amount of flow bypassed annually	•	•	•	•	•	•	•
General Facility	Type and frequency of maintenance	•	•	•	•	•	•	•
	Inlet and outlet dimensions and details	•	•	•	•	•	•	•
Wet Pool	Solar radiation, when available	•		•	•			
	Volume of permanent pool	•		•		•	•	
	Permanent pool surface area	•		•		•	•	
	Littoral zone surface area	•						
	Length of permanent pool	•		•		•	•	

(continued)

Table 3. Continued

Parameter Type	Parameter (1)	Retention (Wet) Pond (2)	Extended Detention Basin (3)	Wetland Pond Basin (4)	Grass/Swale Wetland Channel (5)	Sand/Leaf Compost Filter (6)	Oil & Sand Trap (Vault) (7)	Infiltration and Percolation (8)
Detention Volume	Detention (or surcharge) volume	•	•	•		•	•	•
	Detention basin's surface area	•	•	•		•	•	•
	Length of detention basin		•	•		•	•	•
	Brimfull emptying time	•	•	•		•	•	•
	Half-brimfull emptying time	•	•	•		•	•	•
	Bottom stage volume		•					
	Bottom stage surface area		•					
Pre-Treatment	Forebay volume	•	•	•		•	•	•
	Forebay length	•	•	•		•	•	•
	Other BMPs upstream?	•	•	•	•	•	•	•
Wetland Plant	Wetland type, rock filter present?			•	•			
	Percent of wetland surface at $P_{0.3}$ and $P_{0.6}$ depths			•	•			
	Meadow wetland surface area			•	•			
	Plant species and age of facility	•	•	•	•			

Adapted from Urbonas (1995)

Some researchers have suggested that one should utilize an efficiency measure based upon storm pollutant loads into and out of the BMP on a storm-by-storm basis. This would weight the effectiveness considering that all storms are "equal" in computing the average removal. However, it is readily apparent that all storm volumes and their associated concentrations are not equal. Similarly one could utilize concentrations on a storm-by-storm basis.

One factor that complicates the estimation of effectiveness is that, for wet ponds and wetlands (and other BMPs where there is a permanent pool), comparing effectiveness on a storm-by-storm basis neglects the fact that the outflow being measured may have a limited or no relationship to the inflow. In analysis of rain gauges utilizing SYNOP (Driscoll, et al., 1989), if a basin sized to have a permanent pool equal to the average storm, about 60 to 70% of the storms would be less than this volume. In many cases, the flows leaving may have little or no contribution to flows entering the pond. Therefore, storm-to-storm comparisons are probably not valid. In cases like this, it is probably more appropriate to utilize statistical characterizations of the inflow and outflow concentrations to evaluate effectiveness or, if enough samples are collected (i.e., almost all storms monitored), to utilize total loads into and out of the BMP.

Using the same set of data, Table 4 compares three of the methods including percent removal by storm with a statistical characterization of inflow/outflow concentration and a simple comparison of total loads in and out for the sampled storms. As one can see, the removals estimated differ by up to 19 percentage points. In this record, there are several storm events where inflow concentrations were relatively low and therefore the system was not "effective".

Based upon these factors, it is recommended that the statistical characterization of inflows vs. outflows be utilized (ER). This enhances the ability to conduct statistical tests of whether the reported differences are greater than zero. If enough data on storms are collected (e.g. continuous samples over an extended period), the total loads in and out (SOL) is probably an acceptable method also.

BMP Evaluations – Statistical Considerations

As noted in many studies of urban runoff, the variability in runoff concentrations from event to event is large. If one were to attempt to statistically characterize a BMP influent concentration (and outflow), the more data the better, Figure 7 is a schematic of how more data can improve (reduce confidence interval of) results. As mentioned above, there are a number of types of BMP evaluations that can be conducted. First, the standard evaluation of a single BMP, testing input and output; second, the evaluation of multiple BMPs within a basin (before/after or control ba-

Table 4. Example Wetland TSS Removal

		Concentration		Load			
	Volume of	In	Out	In	Out		% Removal
Storm	(ft^3) Inflow = Outflow	(mg/L)		(lbs)			by storm
1	445,300	352	24	9780	670		93%
2	649,800	30	25	1220	1010		17
3	456,100	99	83	2820	2360		16
4	348,111	433	141	9410	3060		67
5	730,261	115	63	5240	2870		45
	Med	139	65			A	
	Cov	1.48	.86	28,470	9,970	V	48%
	Mean	249	85			G	
		Conc	66%	Loads	65%		

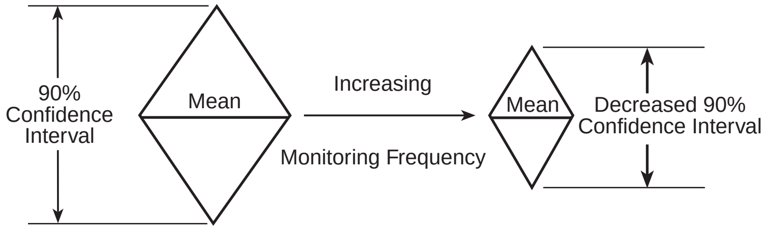


Figure 7. Expected change of 90% confidence interval of station mean with additional data.

sin); and finally a third, the evaluation of a BMP with multiple inlets (where it might be very difficult (expensive) to evaluate the BMP utilizing input/output). All methods should require that a rigorous statistical approach be applied in selecting the number of samples to be collected to assure detection of a given level of change.

As an example of the number of samples required to detect a “true” difference, Table 5 presents an analysis of two of the Portland NPDES monitoring stations (WCC, 1993a) where 10 flow-weighted composited samples were collected. The Fanno Creek station is a large (about 1,200 acres) residential catchment, while the M1 station is a smaller (about 100 acres) mixed land use station. An analysis of a variance-based test was utilized with the existing data to determine how many samples are estimated to be needed to detect a 5%, 20%, and 50% change in the mean concentration at the station. The test was performed considering an 80% probability that the difference will be found to be significant, with a 5% level of significance (Sokal and Rohlf, 1969). This analysis does not consider potential seasonal effects on the collection of data as a factor. Even so, quite a large number of samples would be required to detect a 5% to 20% difference in concentrations. Figure 8 shows a map of the US plotting the average number of storms per year (over 0.1”) as determined by EPA (Driscoll et al., 1989) occur. One can see that in many locations, it would take a number of years of sampling all storm events to be able to detect small differences.

There are numerous examples in the literature where small differences (2 to 5%) are reported based upon much fewer samples than indicated by this analysis. This highlights the need to be more rigorous with regard to statistical testing of reported effectiveness estimates. To detect larger changes, the number of samples becomes reasonable. The mixed land use catchment in Portland is currently being studied for the effectiveness of the implementation of a number of source controls and other controls that do not lend themselves to input/output testing. Examples include maintenance changes (catch basin cleaning, street sweeping), education (business and residences), tree planting, etc. Post-BMP monitoring will be conducted along with qualitative evaluations.

As an example that demonstrates how one could evaluate whether one catchment is different than another, Figure 9 presents results of analysis of stormwater monitoring data collected in Oregon. The figure presents a statistical characterization of land use data, demonstrating that for Total Copper, the open and residential land use stations are statistically different from all other land uses as well as from each other. A similar analysis technique should be employed for all before and after tests, as well as “control” tests.

Table 5. Analysis of Sample Sizes Needed to Statistically Detect Changes in Mean Pollutant Concentrations from 2 Stations in Portland, OR.

		Number of Samples Required to Detect the Indicated % Reduction in Site Mean Concentration*		
Monitoring Site	Parameter	5%	20%	50%
R1 - Fanno Creek Residential	TSS	202	14	4
	Copper	442	29	6
	Phosphorus	244	16	4
M1 - NE 122nd Columbia Slough Mixed Use	TSS	61	5	2
	Copper	226	15	4
	Phosphorus	105	8	3

*80% certain of detecting the indicated % reduction in mean of the EMCs.

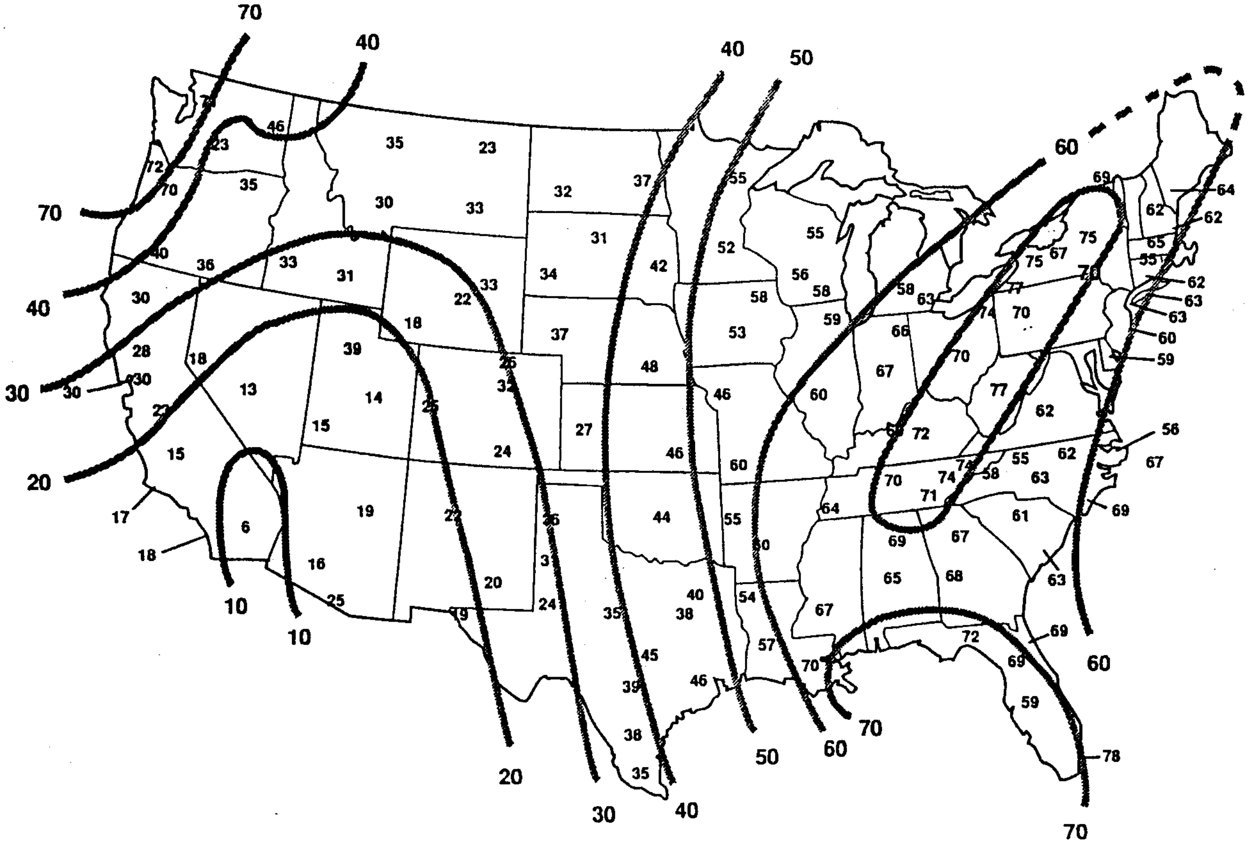


Figure 8. Annual average number of storms. (storms/year)

Other Considerations

There is a need to conduct dry weather analyses between storms on BMPs with dry weather flows; it may be that pollutants captured during storms are slowly released during dry weather discharges.

Biological and downstream physical habitat assessments such as aquatic invertebrate sampling and habitat classification should be explored as an alternative to merely utilizing chemical measures of effectiveness (see Maxted, these proceedings); long-term trends in receiving water quality, coupled with biological assessments, would likely be a much better gauge of the success of the implementation of BMPs, especially on an area-wide basis.

Summary and Recommendations

There is a great need for consistency in the constituents and methods utilized for assessing BMP effectiveness. This

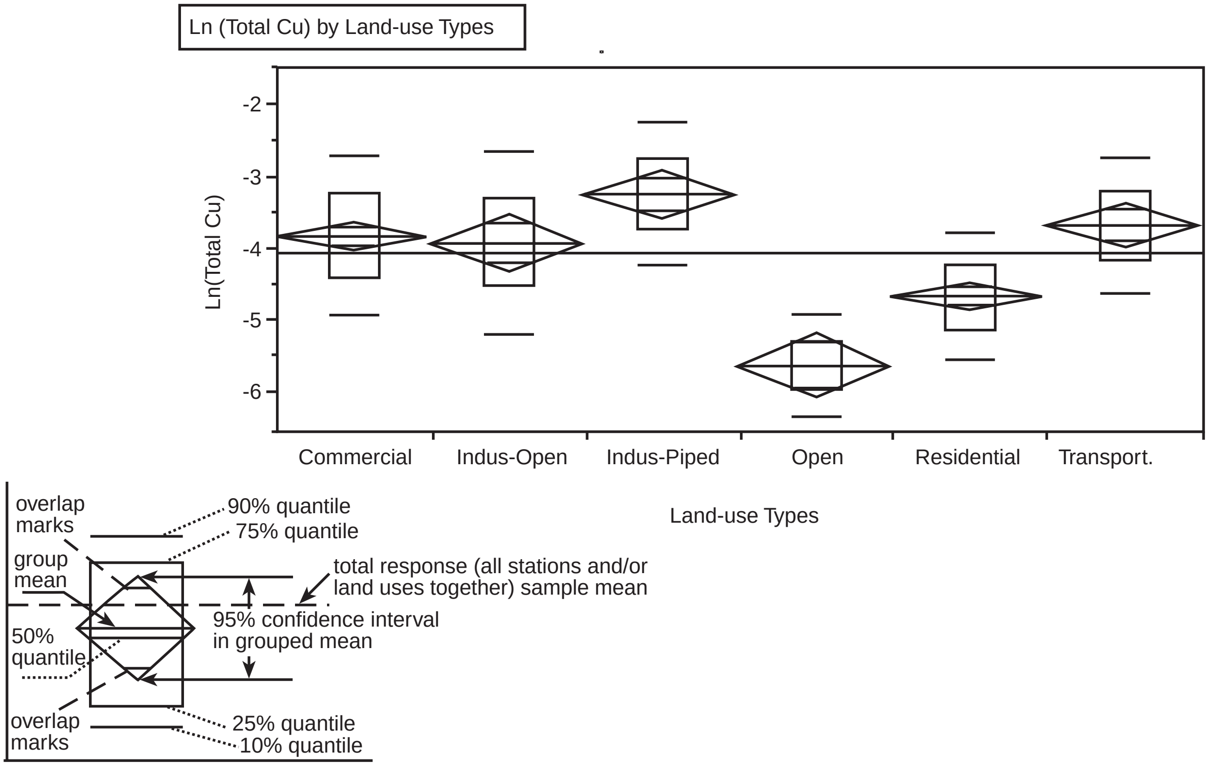


Figure 9. Box Plots of land-use event mean concentration for Oregon stormwater collected for the Municipal Stormwater NPDES Program. (Strecker, et al., 1997)

paper has presented only some of the consistency issues. It is recommended that researchers who undertake BMP effectiveness studies consider the recommendations suggested here, and by Urbonas (1995). It is the authors' opinion that EPA should require studies receiving federal funding to conduct BMP effectiveness studies which utilize standard methods (as suggested here) together with (still much needed) detailed guidance on data collection and sampling methods to improve data transferability.

Acknowledgments

The author wishes to thank Wayne Huber, Gene Driscoll, Gail Boyd, Ben Urbonas, Jonathan Jones, Tom Schueler, Marcus Quigley, and Krista Reininga for their helpful discussions and comments on the subject.

References

Brush, S., M. Jennings, P. J. Young, and H. McWreath. 1994. NPDES Monitoring - Dallas-Fort Worth, TX Area.

Boyd, G.B. 1994. Personal Communication.

Claytor, R. and W. Brown. 1996. Environmental Indicators to Assess the Effectiveness of Municipal and Industrial Stormwater Control Programs. Center for Watershed Protection.

Cooke, T. and C. Lee. 1993. Toxicity Identification Evaluations in San Francisco Bay Area Runoff. Proceedings of 66th Water Environment Federation Conference. Anaheim, CA.

Driscoll, E.D., P.E. Shelley, and E.W. Strecker. 1990. Pollutant Loadings and Impacts from Stormwater Runoff, Volume III: Analytical Investigation and Research Report. FHWA-RD-88-008, Federal Highway Administration.

Driscoll, E., G. Palhegyi, E. Strecker, and P. Shelley. 1989. Analysis of Storm Event Characteristics for Selected Rainfall Gauges Throughout the United States. Draft Report. Prepared by Woodward-Clyde for the U.S. Environmental Protection Agency. 43pp.

Dunne, T. and L.B. Leopold. 1978. Water in Environmental Planning, W.H. Freeman and Company. San Francisco, CA.

Martin, E.H. and J.L. Smoot. 1986. Constituent-Load Changes in Urban Stormwater Runoff Routed Through a Detention Pond-Wetland System in Central Florida. U.S. Geological Survey Water Resources Investigation Report 85-4310.

Maxted, J.R. and E. Shaver. 1999. The Use of Retention Basins to Mitigate Stormwater Impacts to Aquatic Life.

Santa Clara Valley Nonpoint Source Pollution Control Program. 1993. Annual Report, Volume III, Annual Monitoring.

Sartor, J.D., and Boyd, G.B. 1972. Water Pollution Aspects of Street Surface Contaminants. EPA-R2-72-081, U.S. Environmental Protection Agency, Washington D.C.

Sokal, R.R. and F. James Rohlf. 1969. Biometry: The Principles and Practice of Statistics in Biological Research. W. H. Freeman and Company. San Francisco, CA.

Stenstrom, M.K., Silverman, G.S. and T.A. Bursztynsky. 1984. "Oil and Grease in Urban Stormwaters," Journal of Environmental Engineering Division, ASCE. Vol. 110, No.1, pp. 58-72.

Strecker, E. 1994. Constituents and Methods for Assessing BMPs. Proceedings of the Engineering Foundation Conference on Stormwater Related Monitoring Needs. ASCE. Aug. 7-12, Crested Butte, CO.

Strecker, E.W., Kersnar, J.M., Driscoll, E.D., and R.R. Horner. 1992. The Use of Wetlands for Controlling Stormwater Pollution. The Terrene Institute. Washington, D.C.

Strecker, E. and B. Urbonas. 1995. Monitoring of Best Management Practices. Proceedings of the 22nd Annual Water Reources Planning and Management Division Conference, ASCE, NY. pp48-51.

Strecker, E., M. Iannelli, and B. Wu. 1997. Analysis of Oregon Urban Runoff Water Quality Monitoring Data Collected from 1990 to 1996. Prepared by Woodward-Clyde for the Association of Clean Water Agencies.

U.S. Environmental Protection Agency. 1983. Final Report on the National Urban Runoff Program. Water Planning Division, U.S. EPA. Prepared by Woodward-Clyde Consultants.

U.S. Environmental Protection Agency. 1989. Analysis of Storm Events Characteristics for Selected Rainfall Gauges throughout the United States. Draft Report to the U.S. Environmental Protection Agency, Office of Water, Nonpoint Source Division, 43 pp.

U.S. Environmental Protection Agency. 1993a. Manual of CSO Control. EPA-6251R-93-0007.

U.S. Environmental Protection Agency. 1993b. Memorandum. Office of Water Policy and Technical Guidance on Interpretation and Implementation of Aquatic Life Metals Criteria.

Urbonas, B.R. 1994. Parameters to Report with BMP Monitoring Data. Proceedings of the Engineering Foundation Conference on Stormwater Monitoring Related Monitoring Needs. ASCE. August 7-12, Crested Butte, CO.

Urbonas, B.R. 1995. "Recommended Parameters to Report with BMP Monitoring Data. J. Water Resources Planning and Management, ASCE. 121(1), 23-34.

Waschbusch, R. and D. Owens. 1998. Comparison of Flow Estimation Methodologies in Storm Sewers. Prepared by USGS for the Federal Highway Administration. January 16.

Woodward-Clyde Consultants. 1993a. Final Data Report: Data from Storm Monitored between May 1991 and January 1993. Submitted to Bureau of Environmental Services, City of Portland, OR.

Woodward-Clyde Consultants. 1993b. Draft Data Report: Data from Five Storms Monitored between September 1992 and March 1993. Submitted to Department of Public Works, City of Eugene, OR.