Chapter 7
BMP PERFORMANCE ANALYSIS

The analysis of Best Management Practices (BMPs) performance data is often complex and challenging. While the data collection and analysis methods used must suit the objectives of the study, consistency is needed to efficiently facilitate the use of individual study results as well as make broader conclusions with regard to the performance of larger classes of BMPs. Stormwater BMPs have been increasingly implemented over the last few decades to control stormwater runoff and potential pollutant discharge. Recently, Low Impact Development (LID) techniques have become a popular type of stormwater BMP to reduce runoff volumes and associated pollutant loads at or near the source. However, unlike traditional BMPs such as extended detention basins, LID techniques are distributed over a wide area and often do not have a clearly defined inlet and/or outlet. This presents numerous challenges not only for data collection, as discussed previously, but also for data analysis as discussed here.

7.1 Concentrations, Loads, and Volume Reductions

A variety of metrics or measures are available for assessing and quantifying the amount of pollutant conveyed to and from a BMP. Three primary measures are commonly used: concentration of stormwater at an instantaneous point in time (grab samples); the total contaminant load conveyed over a specified duration (e.g., individual storm, daily, weekly); or the event mean concentration (EMC).

7.1.1 Concentrations

Concentrations measured at individual points in time can be useful for BMP efficiency evaluations. Concentrations resulting from samples collected at specific times during an event allow for the generation of a pollutograph. A pollutograph is a plot of the concentration of pollutants as a function of time. Generating pollutographs facilitates the analysis of intra-event temporal variations in runoff concentration. For example, pollutographs can be used to determine if the “first-flush” phenomenon was observed for a specific event. Detailed concentration data is one of the approaches for assessing concentrations of pollutants that have acutely toxic effects, particularly where runoff from storm events constitutes a significant proportion of downstream flow. Under some circumstances, reduction of peak effluent concentrations may be more important than EMC reduction.

The cost of implementing a monitoring program that collects sufficient data to evaluate the temporal variation in runoff and BMP effluent concentration can be high. The trade-off between collecting data from a larger number of events as opposed to collecting detailed concentration data from intra-storm periods often limits the utility of studies that collect detailed concentration data. This type of detailed monitoring is best when focusing on outflow monitoring rather than inflow and outflow.

7.1.2 Contaminant Loads

Contaminant loads are typically calculated by using an average concentration multiplied by the total volume of flow over the averaging period. A variety of methods are available for estimating both the average concentration and the total flow volume. The method chosen depends on the sampling and flow measurement techniques used at the site. Average concentrations may be estimated by collecting time weighted samples, flow weighted samples, or a combination of the two. Likewise, flow data can be collected continuously, intermittently, or modeled from other hydrologic information, such as rain gage information or flow-monitoring conducted in a nearby watershed that has been correlated to the flow at the water quality sampling location. Many BMP monitoring studies focus efforts on water quality sample collection and neglect flow measurement. Accurate flow measurement or well-calibrated flow modeling is essential for load estimation.

Contaminant loads are often most useful when assessing the impact to receiving waters, such as lakes or estuaries, where long-term loadings can cause water quality problems outside of discrete storm events. When the effluent flow rate from a particular BMP is small compared to the flow rate of the receiving water body, potential downstream impairments depend on the absolute load of pollutant rather than the concentrations. For example, loads and load reductions are the central issue in BMP studies that have direct links to receiving water bodies that are regulated under the Total Maximum Daily Load (TMDL) program, particularly when concerned about pollutants being deposited in slow moving systems.

Dry weather flows can contribute substantially to long-term loading. In addition, “on-line” BMPs (e.g., ponds and possibly filters) that have appreciable dry weather flows passing through them can have a reduced “capacity” for storage of wet weather pollutants. For example, pond performance may be affected by the amount of water in the pond before the event, and filters may lose some of their adsorption capacity because of pollutants and other constituents present during dry weather flows.

7.1.3 Event Mean Concentrations

The term “EMC” is a statistical parameter used to represent the flow-proportional average concentration of a given parameter during a storm event. It is defined as the total constituent mass divided by the total runoff volume. When combined with flow measurement data, the EMC can be used to estimate the pollutant loading from a given storm. The primary aim of using the EMC approach to understanding BMP efficiency is to analyze wet weather flows at a site. In most circumstances, the EMC approach provides the most useful means to quantify the pollution level resulting from a runoff event. The collection of EMC data is the primary focus of the BMP Database Project.

7.1.4 Volume Reductions

Since runoff volume reduction is directly associated with contaminant load reductions, it is a key metric used to quantify the performance of any stormwater BMP, particularly BMPs that are designed specifically to store, infiltrate, and evapotranspire captured stormwater. Accurately quantifying volume reductions requires measurements of both inflow and outflow

volumes from the BMP. These volumes are typically based on continuous flow measurements over a specified duration. For BMPs without well-defined inlets or outlets, collection of paired flow measurements may not be practicable, or they may need to be approximated through indirect means such as water level measurements. In some situations, model simulations may be needed to approximate inflow or outflow volumes in order to estimate volume reductions. Refer to Section 8.4.1 for guidance on reaching appropriate conclusions regarding volume reductions through analysis of data or interpreting hydrologic model results.

7.2 Data Analysis and Underlying Principles

The following section provides guidance about the general approaches used to analyze monitoring data and briefly summarizes some of the underlying principles of recommended statistical methods. Due to the overall intent and scope of this manual, many of the methods are merely introduced. Therefore, the reader is encouraged to delve into the statistical literature and data analysis software for the specific details of approaches referenced here.

7.2.1 Descriptive Statistics

The computation of descriptive statistics is a fundamental step in exploratory data analysis. Descriptive statistics include measures of location or central tendency (e.g., mean, median), measures of spread or variability (e.g., standard deviation, interquartile range), and measures of skewness or symmetry (e.g., coefficient of skewness, quartile skew coefficient).

Two general approaches can be taken to compute descriptive statistics: parametric or non-parametric.

Exhibit 7-1 summarizes the parametric and non-parametric statistics commonly used to describe data sets. The parametric and non-parametric methods used to compute these descriptive statistics are briefly described in the sections below.

Exhibit 7-1. Common Parametric and Non-Parametric Descriptive Statistics

Statistic Category	Parametric	Non-Parametric
Measures of Location	Mean	Median
Measures of Spread	Variance, Standard Deviation	Interquartile Range, Median Absolute Deviation
Measures of Skew	Coefficient of Skewness	Quartile Skew Coefficient

Note that a dataset that is highly skewed (e.g., coefficient of skewness <-3 or >3) indicates the data do not arise from a normal distribution and may need to be transformed into a normal distribution prior to applying statistical procedures that depend on an assumption of

normality. Tests of normality should be conducted to verify the transformed data fit the normal distribution.

7.2.1.1 Parametric Statistics

Parametric statistics operate under the assumption that data arise from a single statistical distribution. The specific distribution to which the data are modeled is often chosen by scientific judgment and graphical means, such as the methods described below in Section 7.2.4, and goodness-of-fit tests. The common goodness-of-fit tests are: the *Kolmogorov-Smirnov* (K-S) test and the modified Lilliefors test; the *chi-square* (χ^2) test; the Shapiro-Wilk test; and the probability plot correlation coefficient (PPCC) test.

Once a statistical distribution has been selected, the parameters of the distribution are typically estimated using one of three approaches: (1) method of moments; (2) method of maximum likelihood; or (3) method of L-moments (as described in detail in Chapter 18 of Handbook of Hydrology (Stedinger et al. 1993). While it is beyond the scope of this document to provide specific details about each of these methods, a few qualitative points are provided below:

- Method of moments: This approach utilizes the product moments of a sample to estimate the expected values of the selected distribution parameters (Devore 1995; Dingman 2002). While this method is perhaps the most well known and is the standard for estimating descriptive statistics in spreadsheet statistical software algorithms, it can produce severely biased variance and skewness estimates for small sample sizes and for samples containing a few extreme values. Because of these problems, this method is not recommended for computed descriptive statistics that are to be used to describe the underlying population distribution in order to estimate values at the tails of the distribution (e.g., 90th percentile exceedance probabilities).

- Method of maximum likelihood: Statisticians generally recommend this approach rather than the method of moments approach previously described (Devore 1995). This approach finds parameter values that maximize the joint probability of occurrence for all observed sample values (Chow et al. 1988). As with the method of moments approach, the maximum likelihood estimator can be biased if a sample size is small. However, this method is approximately unbiased when the sample size is large (Devore 1995). The difficulty with this method is that analytical formulae do not exist for every parameter in some distributions (e.g., the shape parameter of the gamma distribution), thus requiring numerical approximations (Chow et al. 1988).

- Method of L-moments: This approach utilizes probability-weighted moments and is nearly as efficient as the method of maximum likelihood, but can be computed analytically using linear combinations of ordered statistics (Hosking and Wallis 1997; Hosking 1990). In addition, the variation and skewness estimates do not suffer from the severe bias problems encountered with the method of moments because L-moments avoid squaring and cubing the data (Dingman 2002).

7.2.1.2 Non-parametric Statistics

Non-parametric statistics are fundamentally based on the ranks[1] of the data with no need to assume an underlying distribution. Non-parametric statistics do not depend on the magnitude of the data and are therefore resistant to the occurrence of a few extreme values (i.e., high or low values relative to other data points do not significantly alter the statistic). The data median is the most basic example of a non-parametric statistic. The median or 50th percentile of a dataset is the value at which half the data lie above and half the data lie below. Depending on the goals of analysis and the uncertainty of the data's underlying statistical distribution, the median may be a more appropriate measure of the central tendency of the data than the sample mean since it is less influenced by the presence of a few outliers. The median EMC may be more representative of the typical or average site storm event discharge concentration because the value is more robust in the presence of outliers, when compared to the mean. The mean EMC for a site, on the other hand, may be completely biased by a single event that had an abnormally high discharge concentration due to an anomalous point source mass release (e.g., a silt fence failing at a construction site).

7.2.2 Relevance of the Lognormal Distribution

The opportune characteristic of water quality data tending toward lognormality is that it can be easily transformed to a normal distribution by simply taking the log of each data point. This transformation allows for parametric statistical procedures that require normality assumptions to be performed. However, a goodness-of-fit test as mentioned above in Section 7.2.1.1 should be performed on the transformed data prior to conducting such parametric procedures.

The lognormal probability distribution is often used to represent environmental data because of the positively skewed nature of the data. The lognormal distribution has been shown to be a good fit for urban stormwater runoff EMC data for many constituents (EPA 1983; Harremoe 1988; Van Buren et al. 1996; Maestre et al. 2005) and this distributional fit has been justified theoretically by Chow (1954). The assumption that a population is lognormally distributed implies that the standard deviation is proportional to the mean and the data are bounded by zero.

The assumption that stormwater data and BMP effluent EMCs are lognormally distributed has been explored in a number of published studies. For example, Van Buren et al. (1996) state that the lognormal distribution may be a better estimate for pond effluent and/or for soluble constituents, an assertion supported by work conducted by Watt et al. (1989). In addition, Maestre et al. (2005) evaluated the probability distributions of the stormwater quality data in the National Stormwater Quality Database. They confirmed that lognormal distributions are very common for the constituents found in that stormwater database, with few exceptions (such as for pH).

[1] In this context, ranks refer to the positions of the data after being sorted by magnitude.

A random variable, x, is said to be lognormally distributed if the distribution of $y = \ln(x)$ is normally distributed with a mean, μ_{ln}, and variance, σ_{ln}^2. The mathematical equation for lognormal distribution is:

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left[-\frac{1}{2}\left(\frac{lnx-\mu}{\sigma}\right)^2\right] \quad x > 0$$ **Equation 7-1**

Where:

μ : sample mean of the untransformed random variable x

σ^2: variance of the untransformed random variable x

x. variable of interest

The lognormal distribution parameters of x are related to the normal parameters of y with the following equations:

$$\mu = \exp(\mu_{ln} + 0.5\sigma_{ln}^2)$$ **Equation 7-2**

$$\sigma^2 = \mu^2\sqrt{\exp(\sigma_{ln}^2) - 1}$$ **Equation 7-3**

A common misconception is that the exponential of the arithmetic mean of the log-transformed variable *y* is the mean of the untransformed variable *x*. However, $exp(\mu_{ln})$ is the geometric mean of the untransformed variable *x*, which is equal to the median of a lognormal random variable, not the mean. Any percentile, p_k, of *x* can be computed using the parameters of *y* as follows:

$$p_k = \exp(\mu_{ln} + z_k\sigma_{ln})$$ **Equation 7-4**

Where:

z_k: *k*th percentiles of the standard normal distribution

Refer to Appendix F for a quick reference equations for transforming between the normal and log-normal distributions.

7.2.3 Comparative Statistics and Hypothesis Tests

The field of comparative data analysis encompasses a series of tests that facilitate determining whether the descriptive statistics of two data sets are significantly different. Such analysis is directly applicable to assessing BMP performance. These methods are capable of comparing totally independent (non-paired) sets of data, such as the effluent concentrations from two different BMP studies, or dependent (paired or matched) data sets, such as the inflow and outflow concentrations from a single BMP. If inflow and outflow data appear to follow different or unknown distributions (e.g., normal or lognormal), or if either data set contains a high proportion (i.e., >15 percent) of non-detects, non-parametric tests may be more appropriate than parametric tests. Both parametric and non-parametric tests are briefly described below.

7.2.3.1 Independent Data Sets

Independent data sets can be compared using the Rank-Sum Test or the *t*-Test. The rank sum test involves joining the data sets and computing the ranks. When dealing with small data sets (n<10, with n being the number of sample points), the functional statistic of this test, W_{rs} is computed by summing all of the ranks of the smaller of the two data sets. For larger data sets, a second statistic, Z_{rs} is computed using W_{rs} , the sample mean, and the sample standard deviation of the combined data sets. A table is then used to assess Z_{rs}. The t-Test can only be used on normally-distributed, uncensored data sets and does not work well for small sample sizes (Helsel and Hirsch 2002). For these reasons, the Rank-Sum Test is often preferred. The difference of magnitudes of two data sets can be quantified using the Hodges-Lehmann test, which is the median of all possible pairwise differences of two data sets. The difference of the sample means is rarely of any value unless the conditions prescribed for the t-Test (uncensored, normally distributed) are met.

7.2.3.2 Paired Data Sets

Matched data sets can be compared using the Sign Test, Rank-Sign Test, and the Paired t-Test. Given two matched data sets, $\boldsymbol{x}$ and $\boldsymbol{y}$, these tests are performed solely on the difference between the two $\boldsymbol{D} = \boldsymbol{x}\text{-}\boldsymbol{y}$. The Sign Test is fully non-parametric, and therefore is often preferred. The number of elements in $\boldsymbol{D}$ that are larger than 0 (noted as S^+) is compared to the number less than 0 (noted as S^-). The Signed Rank Test is used to determine if **x** and **y** are samples of the same population; and if they are, the test is also used to determine whether the difference between the two is only in their location (e.g., median). The paired t-Test is again subject to the assumptions and stipulations associated with other t-Tests mentioned above. The paired t-statistic is computed in the following manner:

$$t_p = \frac{\bar{D}\sqrt{n}}{\sigma}$$ **Equation 7-5**

Where:

$\overline{D}$: mean

σ: standard deviation of the differences

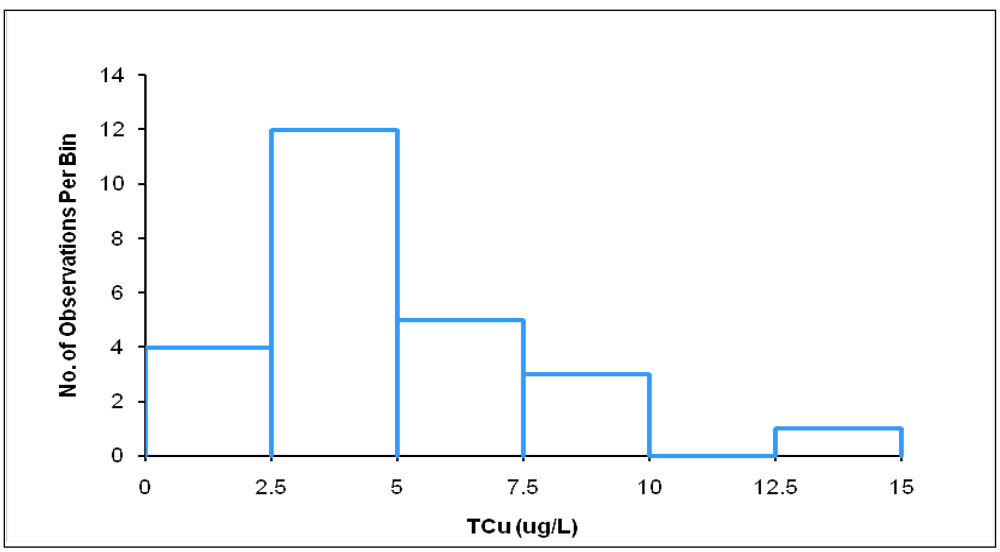
7.2.4 Graphical Data Analysis

Visualizing or graphically displaying data is an essential tool for data analysts. Not only does it provide data analysts with preliminary information about the general characteristics of a dataset, but it also enables them to perform a more comprehensive and statistically valid analysis. Four types of plots are often used to describe and visually display the characteristics of environmental data: histograms; quantile plots; scatter plots; and box plots.

7.2.4.1 Histograms

Histograms are used to visualize the empirical distribution of a single dataset by categorizing the data into bins. The number of data (frequency of occurrence) in each bin is then plotted on the dependent (Y-) axis with the bins themselves on the independent (X-) axis. This practice provides a rough estimate of the shape or symmetry of the probability density function (PDF) of the underlying distribution from which the sample data arise. Exhibit 7-2 is an example of a histogram displaying the frequency of total copper effluent concentrations (TCu) from a wetland basin. Note that the data appear to be skewed to the right (positively skewed) indicating that a lognormal distribution may be an appropriate fit for this data set.

Exhibit 7-2. Example Histogram

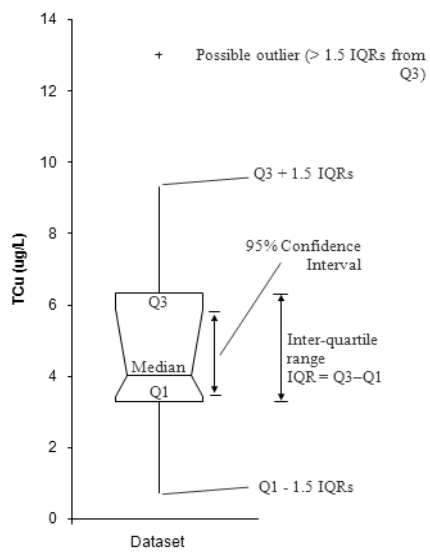


7.2.4.2 Box plots

Box plots (or box and whisker plots) provide a schematic representation of the central tendency and spread of the data. A standard boxplot consists of two boxes and two lines.

The lower box expresses the range of data from the 25th percentile (1st quartile or Q1) to the median of the data (50th percentile, 2nd quartile, Q2). An upper box represents the spread of the data from the median to the 75th percentile (3rd quartile or Q3). The total height of the two boxes is known as the interquartile range (Q3 – Q1). A "step" is 1.5 times the interquartile range. Two lines are drawn from the lower and upper bounds of the boxes to the minimum and maximum data points (respectively) within one step of the limits of the box. Asterisks or other point symbols are sometimes used to represent outlying data points. Some statistical packages, including stand alone software and third-party spreadsheet extensions, also include the confidence interval about the median as notches in the boxes about the center line or can be customized to include specific data percentiles (e.g., 5th, 10th, 90th, and 95th). Exhibit 7-3 shows a boxplot with each characteristic visually displayed.

Exhibit 7-3. Example Boxplot with Definitions



The confidence interval is shown by the location of the notches in the box plot and is based on the work of McGill et al. (1978), which recommends the following definition of the confidence interval:

$$Confidence\ Interval\ of\ Median = Median\ \pm 1.7\left[\frac{1.25IQR}{1.35\sqrt{n}}\right] \qquad \textbf{Equation 7-6}$$

Where:

IQR: interquartile range

n: number of samples

The upper and lower 95 percent confidence limits of the median allow the box plot to be used as a nonparametric, graphical analysis of variance. The extent to which the confidence

intervals for the distributions of event concentrations at the inflow and outflow overlap gives a good indication if the medians can be considered statistically different (i.e., we can reject the null hypothesis that the inflow and outflow medians are the same). In most cases, the Kruskal-Wallis test and the Kolmogorov-Smirnov test support the results of the notched box plot. However, these hypothesis tests are generally more powerful at detecting statistical difference between two sample data sets than simply comparing the confidence intervals about the medians.

7.2.4.3 Quantile Plots and Probability Plots

Quantile plots are used to visually display data for three main reasons: (1) to compare the data distributions of two data sets (called a Q-Q plot); (2) to compare a single data set to a theoretical probability distribution (e.g., normal); or (3) to calculate exceedance frequencies. Quantile plots are constructed by ranking the sample data (i.e., observations) and then calculating the plotting position for each data point. The ranked data are placed on the x-axis and the corresponding plotting positions, or percent less thans (i.e., percentage of total data points below the value on the x-axis), are placed on the y-axis. This produces a sample approximation of the cumulative distribution function (CDF) where the probability of a random sample value being less than or equal to an observation can be directly determined. Conversely, the percent of data points exceeding a water quality threshold (i.e., percent exceedance) can be simply computed as one minus the percentage of data points less than the value on the x-axis.

Depending on the application, there are several different formulas that can be used to compute the plotting position. Helsel and Hirsch (2002) recommend using the Cunnane formula for general use rather than applying a different formula for each application:

$$p = \frac{i-0.4}{N+0.2}$$ **Equation 7-6**

Where:

i: rank of the data point

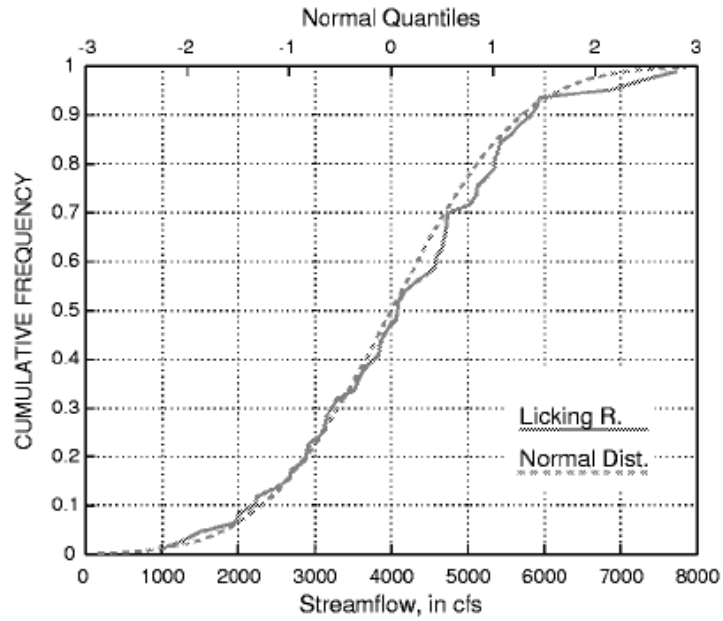
N: number of data points

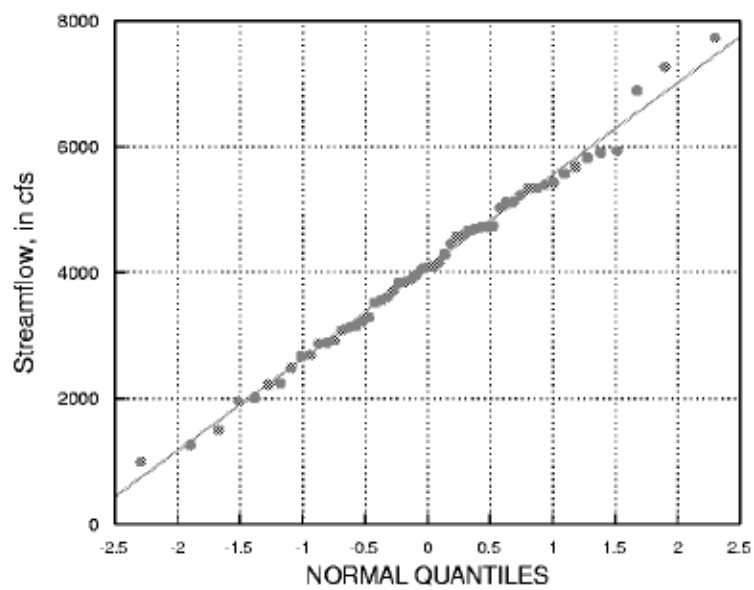
p: plotting position

Probability plots are related to quantile plots, but in this instance the quantiles of the probability distribution, instead of the percent less thans, are plotted against the observations. Both quantile plots and probability plots can be used to determine how well a dataset fits a theoretical distribution. However, rather than plotting the cumulative frequency of the data overlaid with the CDF (or plotting a histogram overlaid with the PDF) a probability plot displays the actual data plotted against quantiles of the probability distribution of interest (e.g., normal quantiles). As such, the agreement of the data with the theoretical straight line is more easily discernable than that of a curved PDF or the CDF. Exhibit 7-4 illustrates a

quantile plot (top) and probability plot (bottom). Basic spreadsheet software can be used to produce quantile plots and probability plots.

Exhibit 7-4. Example Quantile Plot and Probability Plot (Helsel and Hirsch 2002)





Water quality observations do not generally form a straight line on normal probability paper, but they do (at least from about the 10th to 90th percentile level) on log-normal probability plots. This indicates that the samples generally have a log-normal distribution as described previously in this document. That means that many parametric statistical tests can often be used (e.g., analysis of variance), but only after the data is log-transformed. These plots indicate the central tendency (median) of the data, along with their possible distribution type

and variance (the steeper the plot, the smaller the coefficient of variation (COV) and the flatter the slope of the plot, the larger the COV).

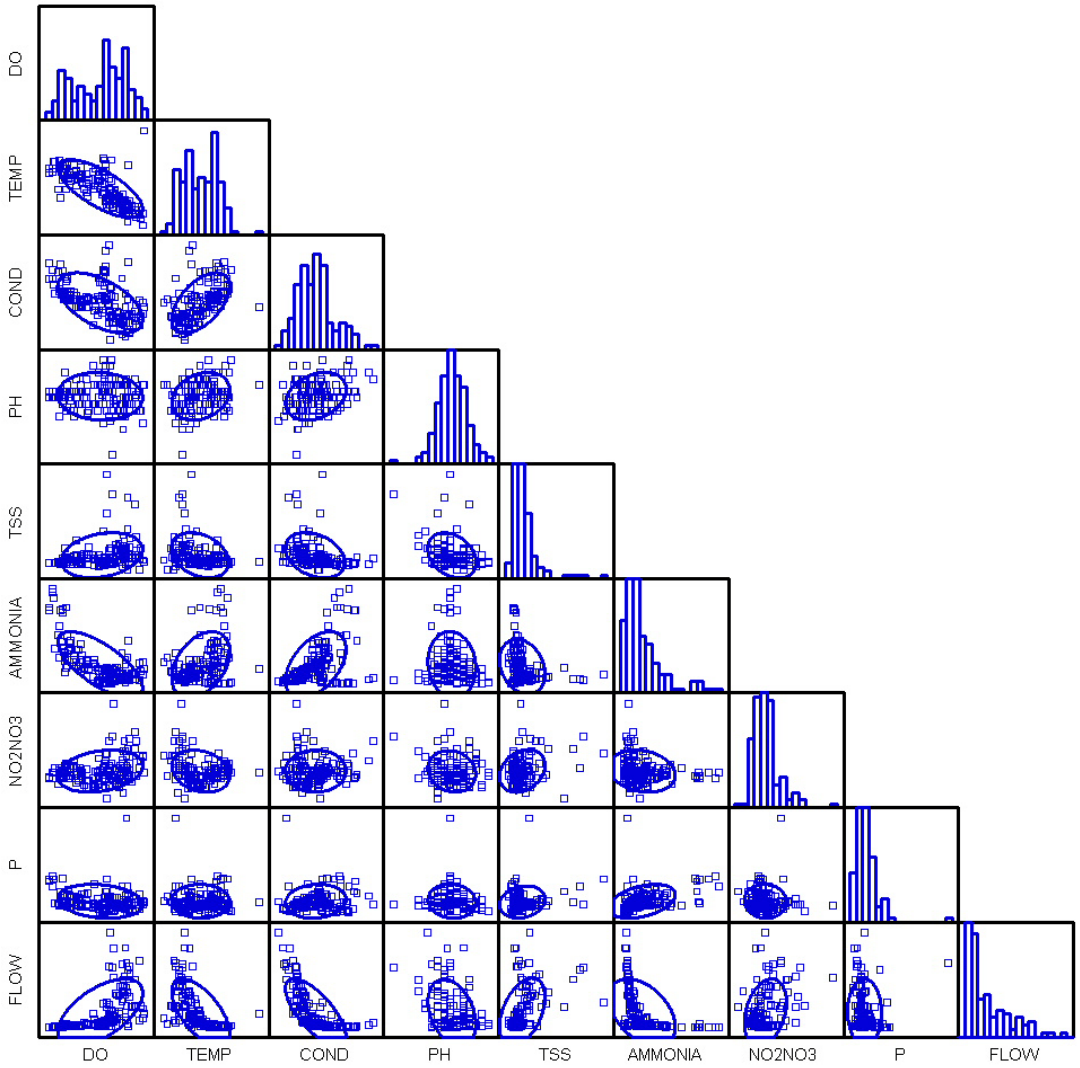
Multiple data sets can also be plotted on the same plot (e.g., different sites, different seasons, different habitats) to indicate obvious similarities or differences in the data sets. Most statistical methods that are used to compare different data sets require that the sets have the same variances, and many require normal distributions. Similar variances are indicated by generally parallel plots of the data on the probability paper, while normal distributions are reflected by data plotted in a straight line on normal probability paper (Burton and Pitt 2001).

Probability plots should be supplemented with standard statistical tests that determine if the data is normally distributed. These tests, at least some available in most software packages, include the Kolmogorov-Smirnov one-sample test, the chi-square goodness of fit test, and the Lilliefors variation of the Kolmogorov-Smirnov test. They are paired tests comparing data points from the best-fitted normal curve to the observed data. The statistical tests may be visualized on a normal probability plot where the best-fit normal curve (a straight line) and the observed data are both plotted. If the observed data crosses the line numerous times, it is much more likely to be normally distributed than if it only crosses the line a small number of times (Burton and Pitt 2001).

7.2.4.4 Scatter Plots

Scatter plots are the most basic of the graphical preliminary investigation tools discussed. These plots are used when discerning a potential relationship between paired data sets or the temporal trend of a single data set. For paired data, the two continuous variables are plotted against each other so that significant patterns or trends become readily apparent upon visual inspection. When looking at potential temporal trends of a single dataset, the independent variable can be time. A scatter plot matrix is useful for evaluating potential relationships between several variables at once and can be a useful starting point for multi-variable regression analysis and when evaluating the potential codependency of variables. Exhibit 7-5 is an example scatter plot matrix.

Exhibit 7-5. Example Scatterplot Matrix of Potentially Related Variables



7.2.5 Analysis of Censored Data

Censored data, or nondetects, include values reported to be only above or below an analytical reporting limit[2]. Nondetects in water quality data are ubiquitous and arise from different

[2] The terms "reporting limit" and "detection limit" are intentionally used loosely and interchangeably in this chapter. While there is clearly a difference between these values—a reporting limit (or quantitation limit) is a threshold based on a measure of the variability or noise inherent in the laboratory process, while a detection limit is a threshold below which measured values are not considered statistically different from a blank signal (Helsel 2005)—it has become commonplace to use either term when referring to nondetects. However, some laboratories will report values between the method detection limit and the laboratory reporting limit. In these

reporting limits based on changes in analytical methods, laboratories, or sample variability. If nondetects are not carefully considered when analyzing data, estimated summary statistics may become biased and nonrepresentative of the monitored site (Helsel 2005).

Four approaches are often utilized to handle nondetects: (1) simple substitution; (2) maximum likelihood estimation (MLE); (3) regression on order statistics (ROS); and (4) Kaplan-Meier (K-M). Each of these methods is briefly described in the sections below.

7.2.5.1 Simple Substitution

Simple substitution replaces all nondetect values with a constant value, such as zero, the detection limit, or half the detection limit. There is no theoretical or mathematical justification for the practice, yet it remains widely used. By substituting constant values, the distribution of the data (i.e., histogram) is altered and the overall variability is reduced. Estimates of the mean and median may become biased high or low depending on the level of censoring and the substitution method employed. It is strongly recommended that simple substitution is avoided, especially when the level of censoring exceeds 5 to 10 percent of observed data. If simple substitution must be performed, the only reasonable value to use is half of the detection limit, as the use of zero or the detection limit can cause more severe bias in computed summary statistics.

7.2.5.2 Maximum Likelihood Estimation (MLE)

With the MLE method, both the censored and uncensored data are assumed to follow a theoretical distribution (as discussed above, the lognormal is often a good choice for water quality data). Summary statistics are then computed as the values that maximize the log-likelihood function (see Helsel 2005 for details). After being first introduced by Hald (1949) and Cohen (1950), maximum likelihood estimators for estimating the statistics of censored data sets have been refined to handle multiple detection limits (Cohen 1976), and several researchers have developed bias-corrected (Cohn 1988; Shumway et al. 2002; Sharma et al. 1995) and robust (Singh and Nocerino 2001; Kroll and Stedinger 1996) MLE formulations.

7.2.5.3 Regression on Order Statistics (ROS)

ROS is a category of robust methods used to estimate descriptive statistics of censored data sets that utilize the normal scores for the order statistics (Shumway et al., 2002). [Normal scores, (also known as Z-scores or z-statistic), are the inverse of the standard normal distribution.] The ROS is a plotting position method developed by Hirsch and Stedinger (1987) and later refined by Helsel and Cohn (1988) for water quality data. In this method, plotting positions are based on conditional probabilities and ranks, where the ranks of the censored (below detection) and uncensored data (above detection) related to each detection limit are ranked independently. After plotting positions for the censored and uncensored

---

cases, Helsel (2005) recommends recensoring the in-between values as <RL and then use a method that can handle multiple detection limits such that the technique accounts for the fact that the RL>DL.

values have been calculated, the log-transformed uncensored values are plotted against the z-statistic corresponding to the plotting position. The best-fit line of the known data points is derived. Using this line and the plotting positions for the uncensored data, the values for the censored data can be extrapolated. The complete "filled in" data set can then be used to estimate descriptive statistics—either by transforming all values back to the original units and computing the statistics (non-parametric formulation) or by computing the statistics in the log-transformed units (parametric formulation) and using lognormal reconversion formulas. Refer to Helsel (2005) or Helsel and Cohn (1988) for details.

7.2.5.4 Kaplan-Meier (K-M)

The K-M method is the standard method for estimating summary statistics for censored survival data (Helsel 2005). It is a completely non-parametric method that utilizes the ranks of the data to estimate "survival probabilities." In the context of water quality data, the survival probability is the probability that a data point would occur below the next incremental concentration given the number of data at or below that concentration or detection limit. Since this method is designed for right-censored data, all observations must be subtracted from an arbitrary value that is higher than the largest observation before it can be used. This transformation results in an empirical cumulative distribution function for the dataset. See Helsel (2005) for details.

7.2.5.5 Recommended Approach for Handling Nondetects

Of the methods described herein, the K-M method is the most robust method for calculating percentiles and works well on both small and large data sets; however, it cannot be used if the level of censoring is greater than 50 percent. Estimates of the mean are biased high using this approach if the lowest reported values are nondetects, which is typically the case. The variance and standard deviation tend to be sensitive to the presence of extreme values in the data set (Helsel 2005). For these reasons, the K-M method is not recommended for general use, particularly when estimates of the mean and its confidence interval are desired.

The MLE and ROS approaches are both useful and can be equally robust and accurate methods for estimating summary statistics. They both require that a distribution be assumed and both have robust and fully parametric formulations. When the distributional assumption is valid, the MLE methods can be more precise, but these methods require larger sample sets ($n \geq 50$) to estimate unbiased summary statistics using the parametric formulation. Probability plotting, as used in the ROS method, is less precise but handles small data sets better. The ROS method is more straight-forward than the MLE, does not require numerical approximations, and can be easily programmed into spreadsheets. As such, the ROS is generally preferred. Many statistical software packages include one or both methods. The ProUCL software package is available free from the Environmental Protection Agency and can be used to compute summary statistics using the ROS method (see http://www.epa.gov/esd/tsc/software.htm).

7.2.6 Bootstrap Methods

Bootstrap methods are a class of data resampling procedures used to estimate summary statistics and their accuracy (standard error). Originally developed by Bradley Efron in 1979, many variations and improvements have been made and the number of applications has grown significantly (Efron and Tibishirani 1993; Chernick 1999). The basic bootstrap method includes sampling from the data set with replacement, calculating the desired descriptive statistics from the sampled data, and repeating several thousand times. The steps of the bootstrap estimation method are described below:

1) Take a sample of size n with replacement (the sampled data point remains in the data set for subsequent sampling) from the existing data set and compute the descriptive statistic, θ_i, from the sampled data. [Singh et al. (1997), recommend *n* be the same size as the original data set.]

2) Repeat Step 1 independently *N* times (e.g., 10,000) each time calculating a new estimate for θ_i.

3) Calculate the bootstrap estimate θ_B by averaging the θ_i's for *i*=1 to *N*

Fundamentally, this bootstrap procedure is based on the Central Limit Theorem, which suggests that even when the underlying population distribution is non-normal, averaging produces a distribution more closely approximated with the normal distribution than the sampled distribution (Devore 1995).

The standard error, $\hat{s}e_b$, of the bootstrap estimates, θ_B, is calculated as:

$$s\hat{e}_B = \left[\sum_{i=1}^{N} \left(\frac{(\theta_i - \theta_B)^2}{(N-1)} \right) \right]^{1/2} \qquad \textbf{Equation 7-7}$$

There are a number of benefits to using the bootstrap method to estimate summary statistics rather than other standard techniques. First, the statistical distribution of the underlying population need not be assumed when using this method. Secondly, the bootstrap method provides more robust estimates of parametric statistics when underlying distribution can be assumed. Lastly, the bootstrap method allows one to compute the accuracy of statistical estimates even when no analytical formula exists (e.g., standard error of the median). Several methods for calculating confidence intervals, or the reliability of an estimate, are also available. Refer to Efron and Tibishirani (1993) for more information.

As with any statistical analysis technique, small data sets can be a problem with the bootstrap. Small data sets underestimate the true variability of the underlying distribution and this underestimation can become magnified with the bootstrap due to multiple repeated values collected during resampling (Chernick 1999). Therefore, as a word of caution, for small data sets (e.g., n<30), the bootstrap may produce inaccurate estimates of population statistics. In these cases, especially when estimating confidence intervals or exceedance

frequencies, parametric methods may be more reliable than the bootstrap, particularly if the data reasonably fit the parametric distribution of interest.

7.3 Error Analysis and Measurement Accuracy

In addition to random uncertainty associated with estimating summary statistics, there are unavoidable sources of error in individual measurements or observations. Sources of measurement error can be associated with the precision of the measuring instrument, the accuracy of the calibration, and the care with which the measurement is made. If the latter two sources of error are minimized or removed, the uncertainty in the measurement is generally on the same order of magnitude as that of the smallest numerical value that can be estimated with the measuring instrument (usually expressed as a percent, or relative error). The true value of that uncertainty typically falls in a range of values that reflect the experimental uncertainty of the measurement. Calculating the mean of multiple measurements (i.e., duplicates) can provide a better estimate of the true value if the measurement errors are random in nature and not systematic.

Problems with instrument precision and/or calibration as well as inaccuracies in the measuring process can lead to indeterminate (random) errors. The magnitude of these indeterminate errors will randomly vary with repeated measurements. There are several ways random errors can be introduced, including operator error, variation in the conditions in which the measuring process is conducted, and the variability of the measuring instrument. QA/QC methods in both the field and laboratory attempt to quantify and control random errors through the use of duplicate samples and blank samples. Laboratory control charts are used to document process results such that adjustments can be made to maintain analytical errors within acceptable limits.

Determinate (systematic) errors have an algebraic sign and magnitude and result from a specific cause introducing the same error into every measurement. Determinate errors are more serious than indeterminate errors because taking the average of multiple measurements cannot reduce their effects. This is because determinate errors have the same sign and magnitude, which prevents positive and negative errors from off-setting each other. Causes of this type of error can include operator bias, (consistent) operator error such as incorrect reading of the instrument, or improper calibration of the measuring instrument.

7.3.1 Expressing Errors

Absolute and relative methods are the standard forms for expressing errors. Absolute error is expressed as a range of values reflecting the uncertainty in the measurement and is reported in the same units as the measurement. Measured values followed by the plus or minus sign express the absolute error.

Relative (or fractional) error is expressed as the ratio of the uncertainty in the measurement to the measurement itself. This is difficult to estimate, because it is a function of the true value of the quantity being measured, which is unknown. Typically this error estimate utilizes the measured value as the “true” value.

The type of measurement and instrumentation can provide an indication of the appropriate form of expressing errors. For example, a pressure probe used to measure depth of flow is likely to have the accuracy of the instrument expressed as a relative percent, while readings on a staff gage would have an absolute error related to the markings on the gage. In these instances the reported depth measurements would be expressed in the same manner as the precision of the measuring instrument.

7.3.2 Propagation of Errors

Quite often, measurements taken of one or more variables are used in equations to calculate the value of other variables. For example, to calculate the area of a rectangle, the length and width are usually measured. To calculate the volume of a cube, the length, width, and height are measured. Each measurement has a potential error associated with it and, as a result, the variable calculated from the combination of individual measurements will also contain some error. The magnitude of the error in the calculated variable can be of a different order than the error associated with any one of the measurements, depending on the algorithm that describes their relationship.

7.4 Performance of BMP and BMP Systems

Previous sections outlined statistical methods to generally analyze water quality data. This section gives guidance on how to apply those methods to evaluate BMP performance.

The efficiency of stormwater BMPs—how well a BMP or BMP system removes pollutants or results in acceptable effluent quality—can be evaluated in a number of ways. An understanding of how BMP monitoring data will be analyzed and evaluated is essential to establishing a useful BMP monitoring study. When analyzing efficiency, it is convenient to classify BMPs according to one of the following four distinct categories:

1) BMPs with well-defined inlets and outlets whose primary treatment depends upon extended detention storage of stormwater (e.g., retention (wet) and detention (dry) ponds, wetland basins, underground vaults).

2) BMPs with well-defined inlets and outlets that do not depend upon significant storage of water (e.g., sand filters, swales, buffers, structural “flow-through” systems).

3) BMPs that do not have a well-defined inlet and/or outlet (e.g., full retention, infiltration, porous pavement, grass swales where inflow is overland flow along the length of the swale).

4) Widely distributed or scattered BMPs where studies of efficiency use reference watersheds to evaluate effectiveness (e.g., catch basin retrofits, education programs, source control programs).

All four of the above categories can also include evaluations to measure BMP efficiency as described below and further discussed in Chapter 8 of this manual.

7.4.1 Comparative Measures of BMP Efficiency

Quantifying the efficiency of BMPs has often centered on examinations and comparisons of "percent removal" defined in a variety of ways. BMPs do not typically function with a uniform percent removal across a wide range of influent water quality concentrations. For example, a BMP that demonstrates a large percent removal under heavily polluted influent conditions may demonstrate poor percent removal where low influent concentrations exist. The decreased efficiency of BMPs receiving low concentration influent has been demonstrated and it has been shown that in some cases there is a minimum concentration achievable through implementation of BMPs for many constituents (Schueler 2000; Minton 2005). Percent removal alone, even where the results are statistically significant, often does not provide a useful assessment of BMP performance.

Exhibit 7-6 provides an overview of the various methods historically used to evaluate BMP performance, and Appendix B provides more detailed discussion of these individual techniques. This chapter focuses on applying the effluent probability approach to evaluate BMP efficiency, which is the approach used by the BMP Database.

Exhibit 7-6. Summary of Historical, Alternative, and Recommended Methods for BMP Water Quality Monitoring Data Analysis

Category	Method Name	Recommendation	Comments	Details
Recommended Method	**Effluent Probability Method**	**Method recommended in this guidance manual.**	**Provides a statistical view of influent and effluent quality.**	Section 7.4.4
Historical Methods	Efficiency Ratio (ER)	Not recommended as a stand-alone assessment of BMP performance. More meaningful when statistical approach is used.	Most commonly used method to date. Most researchers assume this is the meaning of "percent removal." Typical approach does not consider statistical significance of result.	Appendix B1.2
	Summation of Loads (SOL)	Not recommended as a stand-alone assessment of BMP performance. More meaningful when statistical approach is used.	Utilizes total loads over entire study. May be dominated by a small number of large events. Results are typically similar to ER method. Typical approach does not consider statistical significance of result.	Appendix B1.3
Alternative Methods	Percent Removal Exceeding Irreducible Concentration or Relative to WQ Standards/Criteria	Not recommended. May be useful in some circumstances	Typically only applicable for individual events to demonstrate compliance with standards.	Appendix B1.8
	Relative Efficiency	Not recommended. May be useful in some circumstances	Typically only applicable for individual events to demonstrate how well a BMP performs relative to what that BMP is theoretically or empirically able to achieve (as determined from another method).	Appendix B1.8
	Multi-Variate and Non-Linear Models	Possible future use	Additional development of methodology based on more complete data sets than are currently available.	Section 7.4.2

Various historical methods used to evaluate BMP performance are not recommended for use. These methods are as follows:

- Regression of loads (ROL): The assumptions of this method are very rarely valid. This method cannot be universally applied to monitoring data (refer to Appendix B1.4 for more detail).

- Mean Concentration: In this method, it is difficult to "track" the slug of water through BMP without extensive tracer data and hydraulic study. The results are only for one portion of the pollutograph (refer to Appendix B1.5 for more detail).

- Efficiency of Individual Storm Loads: In this method, the storage of pollutants is not taken into account. This method gives equal weight to all storm event efficiencies (refer to Appendix B1.6 for more detail).

- "Lines of Comparative Performance©": This method is not statistically valid due to self-correlation (refer to Appendix B1.8.3 for more detail).

This chapter focuses on applying the effluent probability approach to evaluate BMP efficiency, which is the approach used by the BMP Database.

7.4.2 Multivariate and Non-Linear Model

Reporting efficiency as a percent removal that is calculated based on the difference between influent and effluent concentrations always makes a BMP treating higher strength influents appear to be more efficient than one treating weaker influents, that is if both BMPs are achieving the same effluent quality. A more useful descriptor of efficiency accounts for the fact that weaker influents are more difficult to treat than concentrated ones. A multivariate equation that includes corrections to compensate for this phenomena or a non-linear model may be worth considering for reporting efficiency.

A model that approaches pollutant removal in a manner similar to the reaction rates for complex physical and chemical batch and plug-flow processes may be useful. To date, calibration of such a model for all but the most elementary situations (e.g., settling of solids in relatively simplistic flow regimes) is difficult given the complexity of the real-world problem. As more high quality data become available, other approaches to evaluating BMP efficiency may become apparent.

Currently, effluent quality, as discussed below, is the best indicator of overall BMP performance.

7.4.3 Reference Watershed Methods

Many BMPs do not allow for comparison between inlet and outlet water quality parameters. In addition, it is often difficult or costly to monitor a large number of specific locations if there are many BMPs being installed throughout a watershed (e.g., retrofit of all catch basins). In these cases, a reference watershed is often used to evaluate the effectiveness of a

given BMP or multiple BMPs of the same type. A primary reason to use a reference watershed is to overcome the problem associated with some BMPs that have no clearly defined inlet or outlet point at which to monitor water quality. BMPs with this challenge may include non-structural BMPs, porous pavements, and infiltration practices. The BMP Database allows for a watershed and all its associated data to be identified for use as a reference watershed.

The difficulty in determining the effectiveness of a BMP using a reference watershed approach stems from the large number of variables typically involved. When setting up a BMP monitoring study, it is advantageous to keep the watershed characteristics of the reference watershed and the test watershed as similar as possible. Unfortunately, finding two watersheds that are similar is often quite difficult, and the usefulness of the data can be compromised as a result. In order to determine the effectiveness of a BMP based on a reference watershed, an accurate accounting of the variations between the watersheds, and operational and environmental conditions is needed. The BMP Database explicitly stores some of the key parameters required for normalization of watershed and environmental conditions.

The most obvious parameter used to normalize watershed characteristics is area. If the ratio of land uses and activities within each watershed is identical in both watersheds then the watershed area can be scaled linearly. The loads found at each downstream monitoring station for each event can be scaled linearly with area as well. Difficulty arises when land use in the reference watershed is not found in the same ratio. In this case, either the effects of land use must be ignored or a portion of the load found for each event must be allocated to a land use and then scaled linearly as a function of the area covered by that land use. In many cases, the differences in land use can be ignored (e.g., between parking lots with relatively small but different unpaved areas). The effect of the total impervious area is relevant and should always be reported in monitoring studies. The ratio of the total impervious areas can be used to scale event loads. Scaling the loads based on impervious areas is best used where the majority of pollutants come from runoff from the impervious areas (e.g., parking lots), or where the contaminant of interest results primarily from deposition on impervious surfaces [e.g., total suspended solids (TSS) in a highly urban area]. Methods that attempt to determine BMP performance from poorly matched watersheds yield poor results at best. As the characteristics of the two watersheds diverge, the effect of the BMP is masked by the large number of variables in the system; the noise in the data becomes greater than the signal.

The analysis of BMPs utilizing reference watersheds also requires incorporation of operational details of the system (e.g., frequency of street sweeping, type of device used, device setup). Monitoring studies should always provide the frequency, extent, and other operational parameters for nonstructural BMPs. If the BMP is an alteration of the frequency of a certain practice, the system can be viewed in two ways: (1) as a control/test system; or (2) as a series of data aimed at quantifying the continuous effect of increasing or decreasing BMP frequency. In the first case, the BMP can be analyzed in a manner similar to other BMPs with reference watersheds. In the second case, the loads realized at the monitoring stations need to be correlated with the frequency using some model for the effectiveness of the practice per occurrence.

7.4.4 Effluent Probability Method

The most useful approach to quantify BMP efficiency is to first determine if the BMP is providing treatment (that the influent and effluent mean EMCs are statistically different from one another) and then examine a cumulative distribution function of influent and effluent quality or a standard parallel probability plot.

Before any efficiency plots are generated, appropriate non-parametric (or if applicable parametric) statistical tests should be conducted to determine if any perceived differences in influent and effluent mean EMCs are statistically significant. (The level of significance should be calculated and reported using an appropriate hypothesis test, instead of just noting if the result was significant; assume a 95 percent confidence level.)

Effluent Probability Method is straightforward and directly provides a clear picture of the effluent water quality, which is the ultimate measure of BMP effectiveness. Curves of this type are the single most instructive piece of information that can result from a BMP evaluation study. The authors of this manual strongly recommend that the stormwater industry accept this approach as a standard "rating curve" for BMP evaluation studies.

The most useful approach for examining these curves is to plot the results on a standard parallel probability plot as shown in Exhibits 7-7 through 7-9 below. A normal probability plot should be generated showing the log-transform data of both inflow and outflow EMCs for all storms for the BMP. If the log-transformed data deviates significantly from normality, other transformations can be explored to determine if a better distributional fit exists.

Exhibits 7-7 through 7-9 show three types of results that can be observed when plotting pollutant reduction observations on probability plots. These data were taken from the Monroe Street Wet Detention Pond Study in Madison, WI, collected by the United States Geological Survey (USGS) and the Wisconsin Department of Natural Resources. Exhibit 7-7 for suspended solids (SS) (particulate residue) shows that SS are effectively reduced by the BMP across the entire range of the influent distribution. In contrast, Exhibit 7-8 for total dissolved solids (TDS) (filtered residue) shows poor removal of TDS for nearly the entire range of influent concentrations, as would be expected for a wet detention pond. In this case, the "percent removal" (ER Method) for TDS would be close to zero. Exhibit 7-9, however, shows a wealth of information that is not available from simple statistical numerical summaries. In this plot, filtered chemical oxygen demand (COD) is seen to be poorly removed for low concentrations (less than about 20 mg/L), but the removal increases substantially for higher concentrations. Note that while influent and effluent percentiles are not necessarily paired, the rank order of concentrations was similar for both influent and effluent distributions for all three pollutants (Burton and Pitt 2001).

Exhibit 7-7. Probability Plot for Suspended Solids

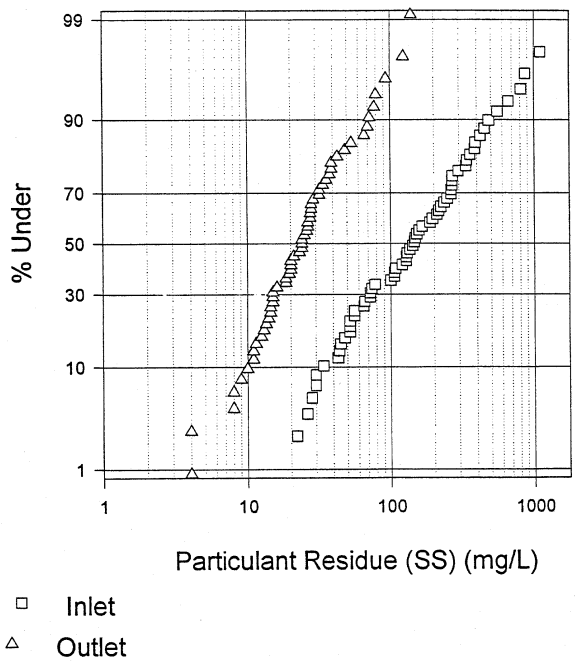


Exhibit 7-8. Probability Plot for Total Dissolved Solids

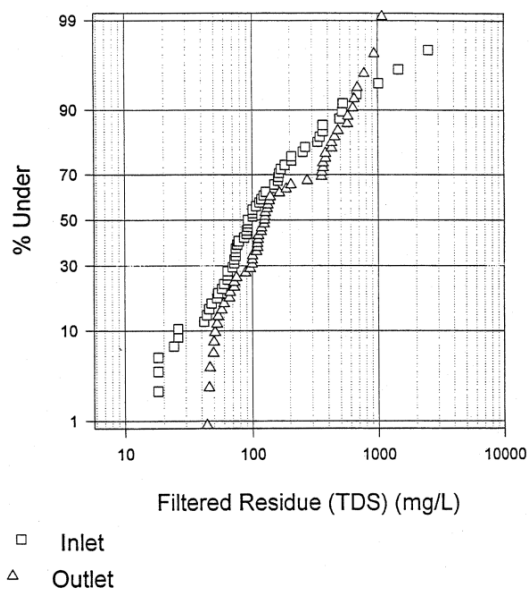
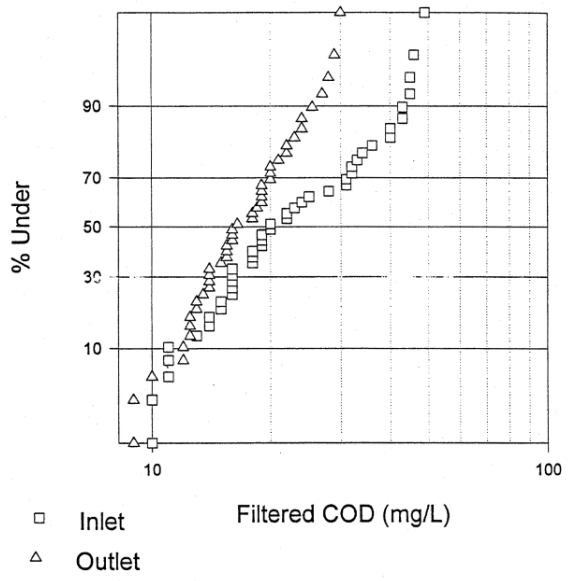


Exhibit 7-9. Probability Plot for Chemical Oxygen Demand



7.4.5 Statistical Measures Employed by the BMP Database

A variety of methods used to evaluate the efficiency and effluent quality of BMPs have already been explored in this Manual. A discussion of data preparation and the statistical and graphical approaches used to evaluate efficiency and effluent quality are presented in this section.

Each BMP study in the BMP Database contains a "Detailed Statistical Analysis Report" (DSAR) for each monitored parameter. The DSAR provides guidance about the efficiency of the treatment practice. The DSAR contains the following elements:

1) Arithmetic estimate of the mean inflow and outflow EMC.

2) Bootstrap estimate of the mean inflow and outflow EMC.

3) Data plots (time series plot; box plot; and probability plot).

4) Summary of distributional characteristics (Shapiro-Wilks W-test and Lilliefors test).

5) Hypothetical test results for non-parametric analysis (Mann-Whitney test).

6) Hypothetical test results for parametric analysis (t-Test on raw and log-transformed data).

7) Test of Equal Variance (Levene Test on raw and log-transformed data).

For the International BMP data sets, it is assumed that both influent and effluent EMCs for most constituents fit well with the log-normal distribution. This assumption appears to be a

good approximation of the distribution of water quality data in most cases. A number of parameters are generally not assumed to fit a lognormal distribution based on analysis of BMP performance data: pH; dissolved oxygen; bacterial counts (e.g., fecal coliform); and turbidity due to the nature of the methods used to quantify these parameters.

For dissolved and particulate water quality constituents, the assumption of log-normality of the data set has been determined by exploring samples of the data in three ways:

1) A comparison between the non-parametric and parametric analyses of variance (p values within 10 percent) has been conducted to assess differences found using the different methods.

2) The Pearson Chi-square test has been conducted and, where there are an appreciable number of data points available, the results of the Pearson Chi-square test indicate that the normal approximation is a good estimate of the central tendency and the distribution of the logarithm of the event concentrations. In cases where a small number of data points are present, there is often very little confidence that either the transformed or non-transformed data are well represented by a normal distribution.

3) Graphical probability plots of influent and effluent have been examined.

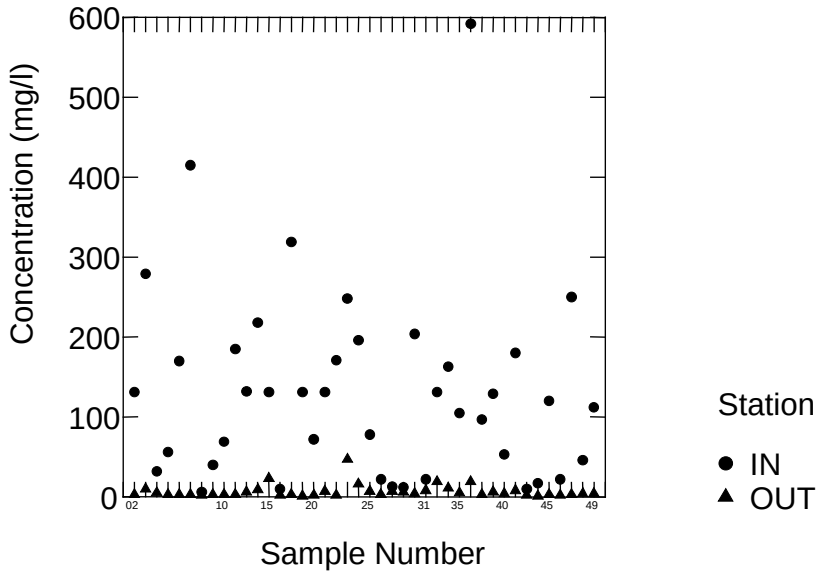
For many summary statistics contained in the BMP Database log-normality was assumed and is typically more appropriate than assuming normality. However, in some specific cases this log-normality may not be appropriate and alternative summary statistics, such as non-parametric statistics, should be chosen. Data plots and goodness-of-fit results should be evaluated before using any particular summary statistics from the BMP Database.

The following provides an overview of the information provided in the periodic statistical summary reports contained in the BMP Database.

7.4.5.1 Influent and Effluent Scatter Plots

Plots showing event concentrations of influent and effluent are included in the DSAR. These plots are compiled based on data collected for each storm. Water quality sample concentrations are identified on a linear scale with inflows and outflows identified using different symbols. These graphs are provided to give an indication of the number of samples collected over the course of the study, which events had paired samples, and the relative difference between influent and effluent concentrations. The sample number indicates each period where samples were collected. This period typically coincides with a storm event; however, this is not always the case. Samples collected during a single dry period are indicated with a separate sample number. If more than one sample is collected at a single location during a period, it is indicated by two of the same symbol at a single sample number. All samples are shown in chronological order. The influent/effluent scatter plot for the Tampa Office Pond (1994 to 1995) for TSS is presented in Exhibit 7-10.

Exhibit 7-10. Linear Influent/Effluent Plot for TSS (Tampa Office Pond 1994-1995)



7.4.5.2 Box Plots

Box plots were discussed as a graphical data analysis tool in Section 7.2.4.2. Box plots are summarized for inflow and outflow concentrations for each BMP study in the Database. A discussion of potential ways of interpreting side-by-side box plots provided in the DSAR is provided below.

There are four primary behaviors observed when comparing distributions of inflow and outflow event concentrations using box plots. These are shown in the following examples:

1) Positive or negative differences where the confidence intervals do not overlap (Exhibit 7-11 and 7-12): This indicates that the median EMCs may not be statistically different. However, if the confidence intervals are nearly overlapping, a more powerful statistical test (e.g., Kruskal-Wallis) should be conducted to determine whether the medians are significantly different.

2) Differences where the confidence intervals appreciably overlap (Exhibit 7-13): In this case, the confidence interval about the median inflow overlaps the confidence interval about the median outflow. The graphical non-parametric analysis of variance (i.e., the notched box plot) indicates that the observed differences in the median are not statistically significant at the 95 percent confidence level.

3) Positive or negative differences where the confidence intervals marginally overlap (Exhibits 7-14 and 7-15): A more powerful statistical test (e.g., Kruskal-Wallis) should be conducted to determine whether the median EMCs are significantly different.

4) Other Cases (Exhibit 7-15): In some cases the 95 percent confidence limit is either in excess of the third quartile or less than the first quartile or both (see Exhibit 7-15). These cases correspond to a distribution of values that is strongly skewed and/or has a low number of samples. The examination of the probability plot may help shed light as to why the confidence limit is outside of the interquartile range.

Exhibit 7-11. Statistically Significant Positive Efficiency as Indicated Through the Box Plot

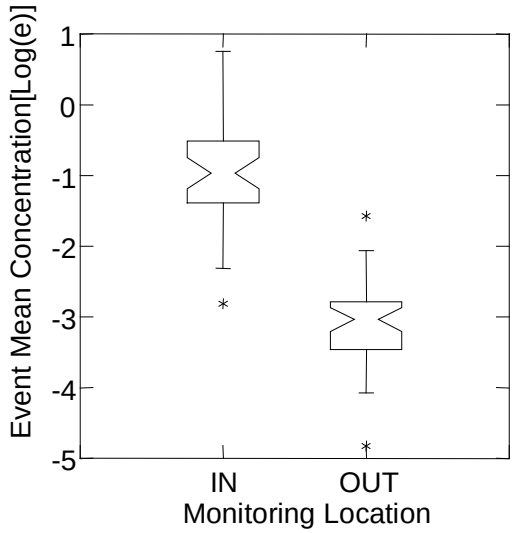


Exhibit 7-12. Statistically Significant Negative Efficiency as Indicated Through the Box Plot

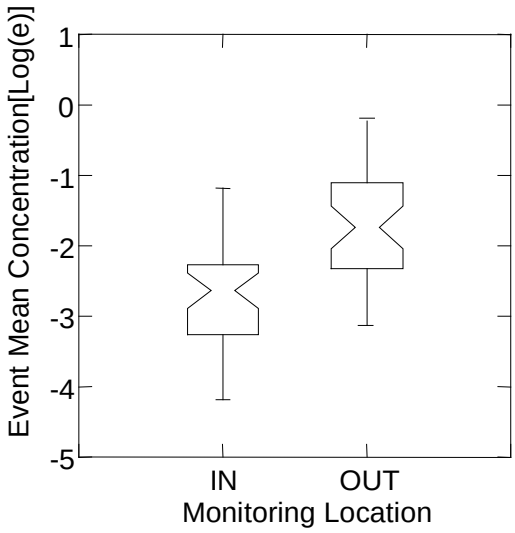


Exhibit 7-13. Statistically Ambiguous Difference in Median Event Concentration as Indicated Through the Box Plot

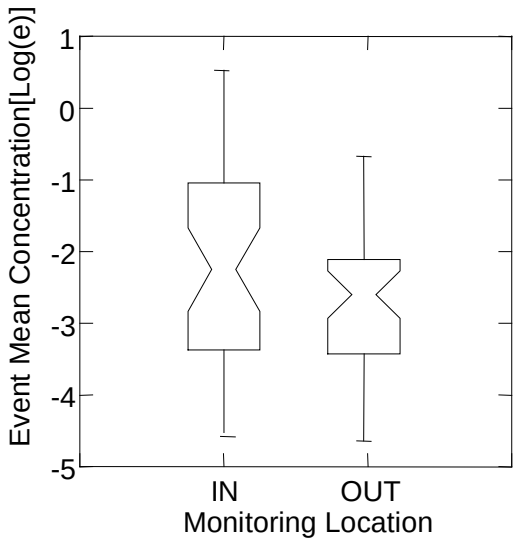


Exhibit 7-14. Marginally Statistically Significant Positive Efficiency as Indicated Through the Box Plot

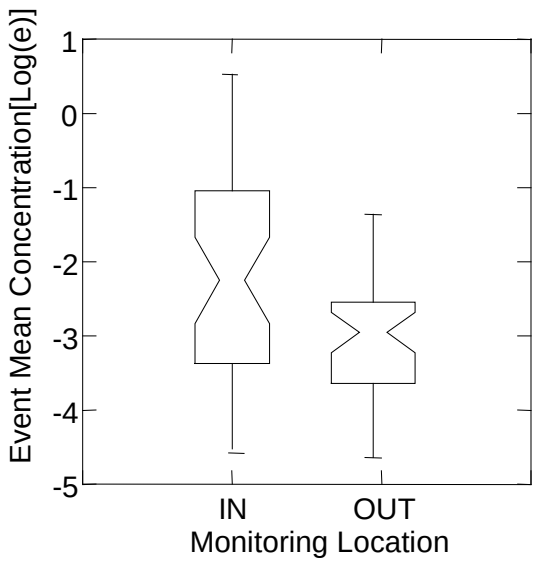
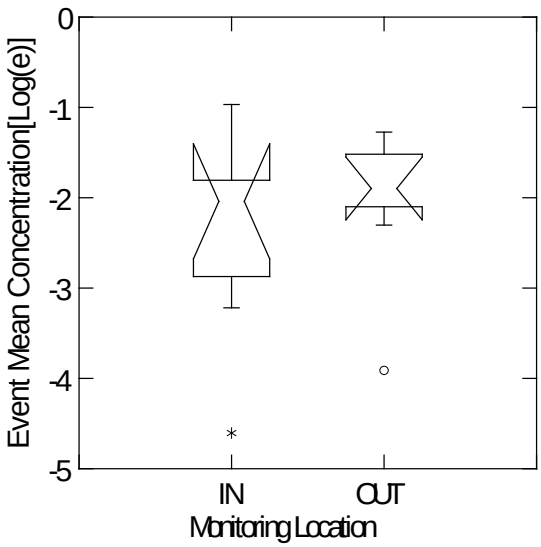


Exhibit 7-15. Example Box Plot Demonstrating Quartiles That Are Inside the Confidence Interval Due to the Small Number of Events Monitored



7.4.5.3 Probability Plots

In accordance with the recommended Effluent Probability Method discussed above, probability plots are included in the DSAR for each BMP and parameter analyzed. Probability plots were chosen for graphical analysis of the water quality concentration data because of the plot's ability to quickly and succinctly relay information about the following:

1) How well data, or transformed data, at each monitoring station are represented by the normal distribution.

2) The mean and standard deviation of the normal distribution and the value of any specific quantile. The slope of the normal approximation is an indication of the magnitude of the standard deviation (straight line); the x-intercept demonstrates the log mean concentration.

3) The relationship between two distributions across the range of quantiles.

4) The presence of any significant outliers.

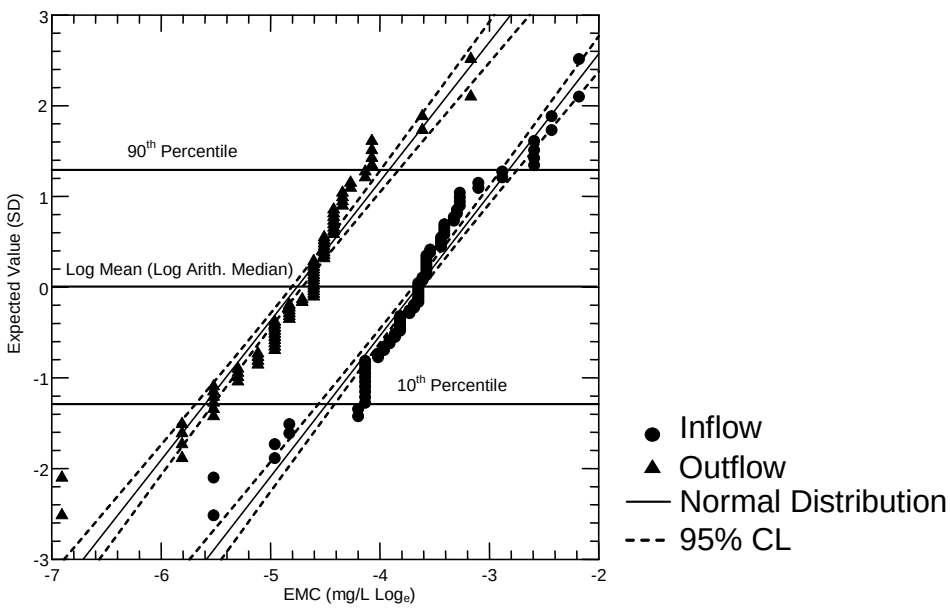
5) The width of the 95 percent confidence interval of the normal approximation.

Note that the quantiles of the inflow and effluent concentrations may not be matched data points as this only occurs if the rank correlation (Spearman's Rho) is precisely 1. For low rank correlations, the 10 percent exceedence concentration in the inflow distribution may not be from the same storm as the 10 percent exceedence concentration in the outflow. Therefore, care must be exercised to not interpret these plots as plots of individual data pairs.

Two sample probability plots are given below in order to explain the range of behaviors that can be encountered when analyzing water quality data. When overlaying normal probability plots for two data sets (typically EMCs from inflow and outflow from a BMP), the results exhibit five primary types of behavior that are described below.

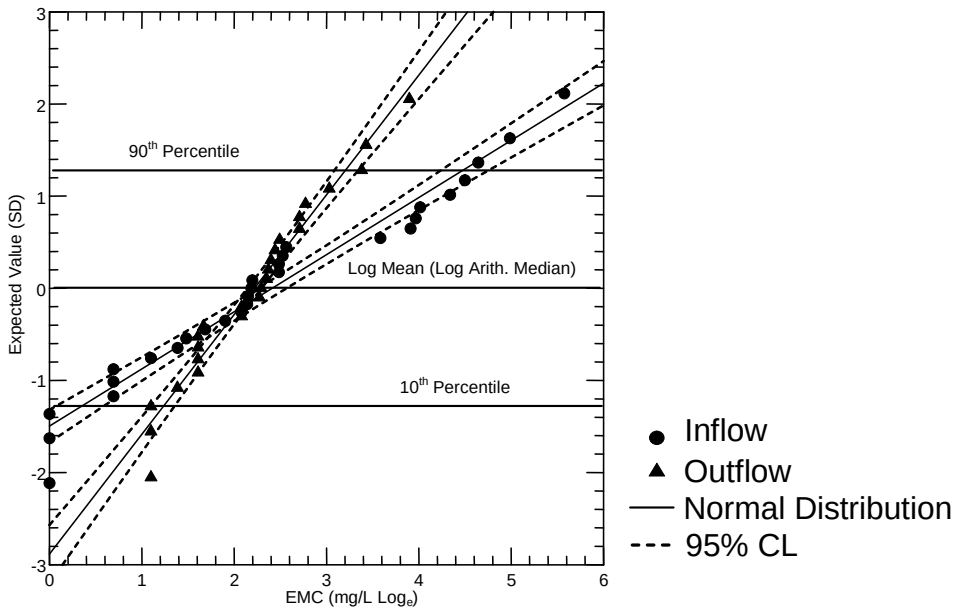
1) The first example (Exhibit 7-16) demonstrates the behavior of two transformed data sets (one from the inflow and one from the outflow of a BMP) that have very similar standard deviations (slope of the normal probability plot) and a uniform difference across the range of quantiles when plotted on a logscale. This indicates that there is a significant difference not only in the log of the mean EMCs, but a significant difference across any given quantile.

Exhibit 7-16. Example Normal Probability Plot for TSS (Tampa Office Pond 1994-1995)



2) The second example (Exhibit 7-17) shows distributions of TSS inflow and outflow EMCs with similar means and different standard deviations. The regression lines cross near the x-intercept, which is the expected value at the mean. This behavior demonstrates negative removal at lower quantiles and positive removal at higher quantiles. This suggests that the BMP may have a minimum effluent concentration that can be achieved, particularly if the intersection point occurs at a low concentration.

Exhibit 7-17. Example Normal Probability Plot for TSS (Tampa Office Pond 1993-1994)



The other three behaviors that are observed when analyzing water quality data are:

1) Water quality data with similar means and similar standard deviations. The difference in the means is not appreciable and the standard deviation of the distributions are similar (no effect from BMP on parameter).

2) Water quality data where the means either have a positive or negative difference and the inflow and outflow distributions overlap at concentrations well below the mean. This shows clear positive or negative removal at the mean but little difference in lower quantiles.

3) Water quality data where the means either have a positive or negative difference and the inflow and outflow distributions overlap at concentrations well above the mean. This shows clear positive or negative removal at the mean but little difference in higher quantiles.

7.4.5.4 Overview of Results of Efficiency and Effluent Quality by BMP Type and Parameter

To summarize the general performance of each BMP type, factsheets have been developed for the BMP Database that are presented as a compilation of effluent concentration statistics and box plots organized by water quality parameter. These plots allow for a quick overview of the range of influent and effluent concentrations across a number of structural BMPs of similar type. Narrative interpretations of the statistical summaries are also included in each fact sheet.

The summaries of each BMP and parameter focus on two separate data analyses:

1) A data set composed of each BMP study's average EMCs over the entire respective monitoring period, grouped by BMP category.

2) A data set comprised of all of the individual effluent EMCs, grouped by BMP category.

For each water quality constituent examined, only those BMP studies reporting at least three influent and effluent EMCs were included in either data set. While this minimum threshold permits the actual calculation of the reported statistics (e.g., mean, median, percentiles), the robustness of such statistics is limited for these smallest samples.

The first data set (averaged EMCs) "weighs" the water quality data for each individual BMP study equally (one average EMC value per BMP study) no matter the number of events monitored, thereby placing the emphasis of the evaluation on whether similar types of BMPs at a variety of different sites achieve comparable average effluent quality. This analysis mutes the influence of individual events, and does not favor BMP studies that report a relatively large number of EMCs. The second analysis compares the distribution of effluent water quality from individual events by BMP category, thereby providing greater weight to those BMPs for which there are a larger number of EMCs reported. This represents an important distinction between the two analyses, and it is essential that interpretation of the performance summaries reflect how the data has been compiled and presented.

Notched box plots are used to graphically display the categorized distributions from both data sets. The notches encompass the 95 percent confidence interval of the median (averaged EMCs or individual EMCs, depending on the analysis) and provide a graphical, nonparametric means of assessing the difference between the central tendencies of multiple distributions. A logarithmic scale was determined to be best suited for plotting the data. The log-scale box plots were created utilizing the following method to calculate the upper and lower confidence levels:

1) The natural logs of the effluent values (averaged EMCs or individual EMCs, depending on the analysis) for a given BMP category are sorted in ascending order.

2) The upper and lower quartiles (i.e., the 75th and 25th percentiles) are calculated, following Tukey (1977).

3) The confidence interval of the median is calculated based on the upper and lower quantiles, following McGill et, al, (1978).

4) The median and confidence interval is translated back to arithmetic space. These values are used to delineate the upper and lower bounds of the notch on the box plots.

For both the distributions of averaged EMCs by BMP category and the distributions of individual EMCs by BMP category, the arithmetic values of the median and associated upper

confidence level (UCL) and lower confidence level (LCL) are provided in the table that accompanies each summary.

An assessment was also made of the difference between the median effluent values and the corresponding influent values for both data sets. This assessment is critical because it provides a measure of whether or not the data indicate a statistically significant difference in pollutant levels between the influent and effluent. To perform this test, the median, UCL, and LCL for influent values were calculated in the same manner as for the effluent. A significant difference between the median influent and effluent values is assumed if their respective confidence intervals do not overlap; otherwise, the difference is not considered statistically significant. The same test may be performed graphically by plotting influent and effluent notched box plots side-by-side and comparing the confidence limits visually.

In many instances, no significant difference between influent and effluent medians was determined. Therefore, it is not possible to determine with any certainty whether the BMP had an effect or simply that the characteristics of the runoff treated (for example, low influent concentrations) govern the distribution of effluent values. Where the analysis of significant difference indicates that effluent levels are *greater* than influent, this is noted in the text and as a footnote to the tabulated values.

A brief synopsis describing the parameter in question and the mean effluent concentration for each BMP is given in the summary page. An example of a BMP parameter summary sheet is given in Exhibit 7-18.

Note that there are several limitations to these generalized BMP performance summaries due to the grouping of data from several studies.

1) The studies grouped into the various BMP categories may have widely varying designs that have direct influence on their performance.

2) A large effort has been made to assure the quality of the data; however some studies may have been miscategorized.

3) Local hydrology and drainage area conditions may have a large influence on the performance of BMPs.

4) There are an unequal number of data points within and among each study and BMP category. This may bias the statistical summary results.

While the BMP performance fact sheets are a major step toward quantifying the general performance of BMPs in the Database, it is recognized that there is more research to be done and more data to collect in order to increase the confidence in the statistical summaries. In the future, there may be enough data to separate individual studies based on a variety of factors that are suspected to influence performance.

7.5 Conclusions

The analysis of Best Management Practices (BMPs) performance data is often complex and challenging. Consistency in performance analysis is needed to efficiently facilitate the use of individual study results as well as make broader conclusions with regard to the performance of larger classes of BMPs. Performance analysis methods must account for changes in concentration, pollutant load, and volume of discharge that result from the BMP. Methods must also account for measurements below detection limit, small samples size, availability of paired or unpaired datasets, and unknown statistical distributions. While a variety of analysis methods can yield meaningful results provided that appropriate practices are followed, the effluent probability method is recommended because of its flexibility and reliability in describing the fundamental aspects of BMP performance. Data from the BMP Database can be visualized and compared via scatter plots, box plots and probability plots in order to reach conclusions about the absolute and relative performance of BMPs.

This guidance document only briefly describes some of statistical methods available to the data analyst and is not intended to be a standalone statistical reference. Users are encouraged to seek additional information on how to apply the methods described in this manual and their limitations from the vast body of statistical literature and available statistical software.

Exhibit 7-18. Example of BMP Parameter Summary Sheet

Analysis of Treatment System Performance - Solids

SEE INTRODUCTION FOR INTERPRETATION OF THESE FIGURES

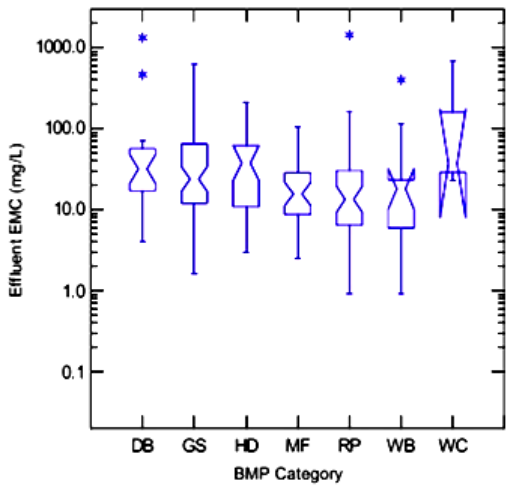


Figure 1. Mean effluent TSS concentration by BMP category

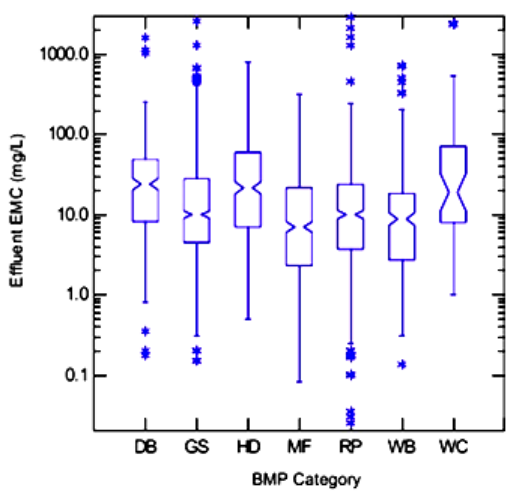


Figure 2. Individual effluent TSS EMCs by BMP category

Total Suspended Solids (mg/L)

Total suspended solids (TSS) represent the most widely reported stormwater constituent in the International Stormwater Best Management Practices (BMP) Database. Information regarding particle size distributions or settling velocities among the studies included in the database is very limited, and no distinction based on these factors is made between BMP studies analyzed. Particle size distribution may play a significant role in BMP performance. For example, coarse sand settles more rapidly than finer particles associated with clayey or silty soils.

Although EPA does not provide a national recommended numeric water quality criterion for TSS, many NPDES construction dewatering and wastewater permits identify 30 mg/L as the average permissible TSS concentration. Median concentrations for all of the BMP categories are below 30 mg/L.

Analysis of Mean Effluent TSS Concentration by BMP Category (one value per BMP Study)

Average effluent TSS concentrations are significantly lower than average influent for biofilters, media filters and retention ponds. Median averaged effluent concentrations for detention basins, biofilters, wetland channels and hydrodynamic devices are above 15 mg/L, while those for media filters, retention ponds and wetland basins range between approximately 10 to 14 mg/L.

Media filters, biofilters and hydrodynamic devices are all primarily flow-through systems (i.e. no significant detention of flows). Of the storage-type categories, those which include some kind of permanent pool (i.e., retention ponds and wetland basins) exhibit significantly lower effluent levels. Hydrodynamic devices that include storage components were not analyzed separately in this summary report.

Analysis of Effluent TSS Concentrations by BMP Category (all individual EMCs included in dataset)

Median effluent TSS EMCs for all BMP categories exhibited statistical significance between influent and effluent EMCs. Effluent concentrations appear to be greater than influent concentrations for wetland channels.

	BMP Category	Number of BMPs	Median of Avg. Effluent (95% Confidence Interval)[1]			Significant Difference Between Average Influent and Effluent[2]	Median of Effluent EMCs (95% Confidence Interval)[1]			Significant Difference Between Influent and Effluent EMCs[2]
			Median	LCL	UCL		Median	LCL	UCL	
DB	Detention Basin	22	31.04	16.07	46.01	NO	25.00	21.26	29.04	YES
GS	Biofilter	56	23.92	15.07	32.78	YES	10.00	9.08	11.02	YES
HD	Hydrodynamic Device	30	37.67	21.28	54.02	NO	21.90	18.49	25.93	YES
MF	Media Filter	33	15.86	9.74	21.98	YES	7.60	6.56	8.81	YES
RP	Retention Pond	43	13.37	7.29	19.45	YES	10.00	8.93	11.20	YES
WB	Wetland Basin	14	17.77	9.26	26.29	NO	9.40	7.85	11.25	YES
WC	Wetland Channel	3	37.25	8.02	187.13	NO	19.00	10.93	33.03	YES[3]

1. Calculation of confidence interval based on McGill et al (1978), from the natural log of the quantiles.
2. Based on non-parametric analysis of difference in median values.
3. Indicates that effluent is significantly greater than influent.

7.6 References

Chernick, M.R. 1999. *Bootstrap Methods – A Practitioner's Guide.* John Wiley & Sons, Inc., New York.

Chow, V.T., Maidment, D.R., and L.W. Mays. 1988. *Applied Hydrology*. McGraw-Hill, Inc. New York, NY.

Cohen, A.C. 1950. *Estimating the Mean and Variance of Normal Populations for Singly Truncated and Doubly Truncated Samples*. Annals of Mathematical Statistics, Volume 21(4). 557-569 p.

Cohn, T.A. 1988. *Adjusted Maximum Likelihood Estimation of the Moments of Lognormal Populations from Type I Censored Samples*. U.S. Geological Survey Open-File Report 88-350.

Devore, J.L. 1995. *Probability and Statistics for Engineering and the Sciences.* Fourth Edition. Brooks/Cole Publishing Co., Pacific Grove, CA.

Dingman, S.L. 2002. *Physical Hydrology*. Second Edition. Waveland Press, Inc., Long Grove, IL.

Efron, B. and Tibishirani. 1993. *An Introduction to the Bootstrap.* Chapman & Hall, New York.

Hald, A. 1949. Maximum *Likelihood Estimation of the Parameters of a Normal Distribution which is Truncated at a Known Point*. Skandinavisk Aktuarietidskirift Volume 32, Issues 1 and 2. 119-134 p.

Helsel, D.R. 2005. *Nondetects and Data Analysis – Statistics for Censored Environmental Data.* John Wiley and Sons, Inc., Hoboken, New Jersey.

Hosking, J.R.M. 1990. L-*Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics*. Journal of Royal Statistical Society, *B*, Volume 52, Issue 2. 105-124 p.

Hosking, J.R.M and J.R. Wallis. 1997. *Regional Frequency Analysis: An Approach Based on L-Moments*. Cambridge University Press, Cambridge, United Kingdom.

Helsel, D.R. and R.M. Hirsch. 2002. *Statistical Methods in Water Resources*. U.S. Geological Survey, Techniques of Water-Resources Investigations Book 4, Chapter A3. U.S. Geological Survey, Water Resources Division, Reston, VA.

Helsel, D.R. and T.A. Cohn. 1988. *Estimation of descriptive statistics for multiply censored water quality data*. *Water Resource Res.,* Volume 24, 1997-2004.

Hirsch, R.M., and J. R. Stedinger. 1987. *Plotting positions for historical floods and their precision*. Water Resource. Res., Volume 23, Issue 4. 715-727 p.

Kroll, C.N. and J.R. Stedinger. 1996. *Estimation of Moments and Quantiles Using Censored Data.* Water Resource Res., Volume 32. 1005-1012 p.

McGill, R., Tukey, J.W., and W.A. Larsen. 1978. *Variations of Box Plots*. The American Statistician, Volume 32, Issue 1. 12-16 p.

Minton, G.R. 2005. *Stormwater Treatment.* Resources Planning Associates, Seattle, WA.

Schueler, T.R. 2000. *Irreducible Pollutant Concentrations Discharged from Stormwater Practices*. Article 65 in The Practice of Watershed Protection, Schueler, T.R. and H.K. Holland (eds.). Center for Watershed Protection, Ellicott City, MD.

Sharma, M., McBean, E.A., and N. Thomson. 1995. *Maximum Likelihood Method for Parameter Estimation in Linear Model with Below-Detection Data*. Journal of Environmental Engineering, Volume 121, Issue 11. 776-784 p.

Shumway, R.H., Azari, R.S., and M. Kayhanian. 2002. *Statistical Approaches to Estimating Mean Water Quality Concentrations with Detection Limits*. Environmental Science. Technology, Volume 36. 3345-3353 p.

Singh, A. and J. Nocerino. 2001. *Robust Estimation of Mean and Variance Using Environmental Data Sets with Below Detection Limit Observations*. U.S. Environmental Protection Agency Clearance No. 01-062.

Stedinger, J.R., Vogel, R.M., and E. Foufoula-Georgiou. 1992. *Handbook of Hydrology, Chapter 18: Frequency of Analysis of Extreme Events*. McGraw Hill, Inc., New York, New York.

Tukey, J.W. 1977. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.