

### III. ANALYSIS AND RESULTS

This report will focus on concentration data of four water quality parameters: total phosphorus, ammonia-nitrogen, nitrate-nitrogen, and fecal coliform bacteria. This introductory section provides a description of each of these parameters, and an overview of the results from the entire WVDA data set from July 1998 through June 2004. Over 13,000 samples were collected at 100 sites throughout this region during this period. All of the sampling sites were in basins that were affected by anthropogenic (human) influences and, therefore, water quality in all sites show the signature of human uses of the land.

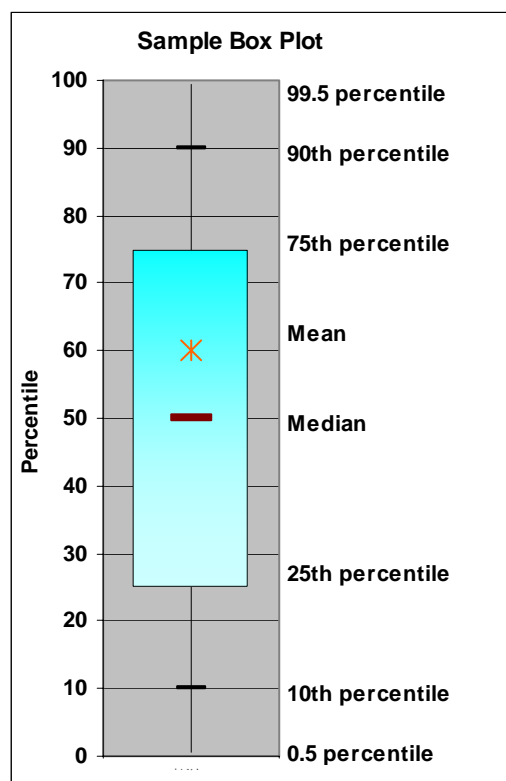
In order to reduce sampling bias, all samples were collected based on a preset schedule, and were collected regardless of weather conditions, including both high and low flows, and during periods of normal, high and low precipitation. A description of the WV Department of Agriculture's field and laboratory methods is provided in Appendix 1.

The following sections will present site-by-site data from selected watersheds, followed by a comparison of selected watersheds with each other; these sections will also show how each site compares to the entire WVDA data set. The data in this report were assessed in three ways: graphical comparisons between sites and watersheds using box plots, trend analysis based on time series regressions and the seasonal Kendall test (analyses by WVU), and analysis of variance. Summary statistics are provided in Appendix 2.

#### Analytical Methods

It is often a challenge to analyze water quality data where the main sources of the parameters being studied are non point, because the data varies tremendously. It is not unusual to have a large number of relatively low values, and a few extremely high values that occur during or following significant precipitation. Box plots (Figure 7 at right) are commonly used for side-by-side visual comparisons of this kind of data. The top and bottom of the box are the 25<sup>th</sup> and 75<sup>th</sup> percentiles, and the area contained within the box is called the interquartile range; the middle 50% of the data is contained by the box, 25% of the data falls below and 25% falls above the bounds of the box. The interquartile range provides important information on how variable the data is. For example, if a sampling location has a relatively constant source of some substance, the interquartile range would be relatively narrow, while if the source was highly variable, the range would be relatively large.

The average, or mean, is often used to indicate the "central tendency" of statistical data. Central tendency is simply where the preponderance of measurements in the data occur. You can think of central tendency as the data's balance point. However, the mean is usually a poor measure of central tendency for water quality data (i.e.: data with a large number of relatively low values and a few extremely high values) because it is pulled (or skewed) away from the balance point by even a few very high values. The median (which equals the 50<sup>th</sup> percentile of the data) is generally the best measure of "central tendency" for water quality data, and is indicated as a brown



**Figure 7. Sample Box Plot.** The brown bar is the median or 50<sup>th</sup> percentile, the orange star is the mean, and the box encloses the middle 50% of the data.

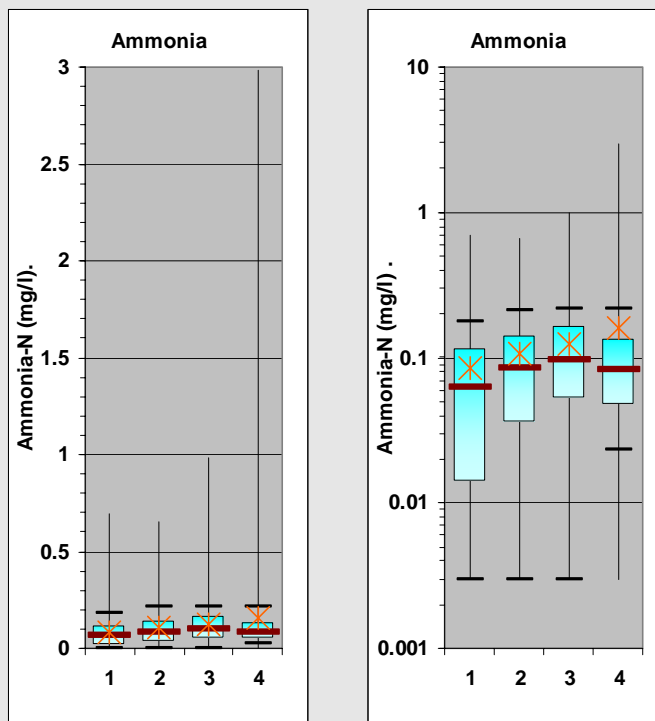
horizontal line in the box. The mean is indicated by an orange star. The distance between the mean and the median is a good indication of how "skewed" the data is.

The 90<sup>th</sup> and 10<sup>th</sup> percentiles are indicated by top and bottom of the horizontal black lines passing through the box. The 99.5<sup>th</sup> and 0.5<sup>th</sup> percentiles are found at the end of the lines that extend above and below the box. Minimum and maximum values were not shown because maximum values are often much higher than the majority in non point source data and displaying them can seriously distort the main purpose of the graphs – a visual comparison of central tendency. The 10<sup>th</sup> and 0.5<sup>th</sup> percentiles, in combination, provide important information on the potential for and frequency of low concentrations at each site. Similarly, the 90<sup>th</sup> and 99.5<sup>th</sup> percentiles, in combination, provide important information on the potential for and frequency of high concentrations at each site, particularly important for substances that come from non point sources that move into streams primarily via overland runoff. Due to the extreme range of the data, logarithmic scales were chosen for display of three of the four parameters: total phosphorus, ammonia-nitrogen, and fecal coliform bacteria. Logarithmic scales, while often necessary for graphing water quality data, distort the appearance of the data and can be very difficult to interpret if you are not used to them. (See box “The Trouble with Log Scales” below)

### The Trouble with Log Scales

Water quality data in watersheds dominated by non point sources often have a large number of relatively low and a few extremely high concentrations. Logarithmic scales are often required to display this kind of data graphically. However, logarithmic scales distort the appearance of the data and must be interpreted with care.

For example, the two graphs in this box display the same ammonia data, with the graph on the left using a normal scale, the graph on the right logarithmic. Because most of the data is clustered at the lower end of the normal scale graph, it is difficult to see differences for statistics in the range from the minimum to the 90<sup>th</sup> percentile, while the upper extremes of the data are clear. Conversely, the distribution of data in the “central” part of the logarithmic graph are displayed clearly, but with graphic “distances” increasingly compressed towards the top of the graph the very high 99.5<sup>th</sup> percentile for sample 4 appears not much different than the other three samples.



Trends in the data were determined by West Virginia University using three methods; time series linear regression with a trend variable, seasonal and modified Kendall Trend Tests, and fitting locally weighted linear regressions (LOWESS) to assess the curvilinear trends in the data. Time series regression was used to determine the relationship between the sampled concentrations and time. Due to varying climate and precipitation throughout a given year, there is a strong potential for seasonality in the data. As such, the seasonal Kendall test was performed on each constituent at each site. This test compares the data on a season-by-season basis to determine whether concentrations are affected by seasonality, and whether trends exist in the data. The body of this report provides simplified results of these analyses,

indicating only if a significant trend was detected and the direction of significant trends. Complete results of the time series and seasonal Kendall tests are provided in tabular form in Appendix 3, along with more technical descriptions of the statistical methods used.

One way analysis of variance (ANOVA) was used to detect differences between watersheds. Because of the severe "skewness" of the data, common in water quality data affected by non point source pollution, ANOVA's were run on rank-average transformed data for comparison of median concentration distributions. An alpha value of 0.05 was used as the threshold for statistical significance. If a significant difference among group medians was detected, Tukey's multiple comparison test was used on the rank transformed data to determine where differences were located. This method is described by Helsel and Hirsh (1992) and is commonly used for water quality data analysis. ANOVA analysis was conducted using JMP Statistical Discovery Software (version 4.0.2) (SAS Institute, 2000).