THESIS

THE ROLE OF DATA ANALYSIS METHODS SELECTION AND DOCUMENTATION IN PRODUCING COMPARABLE INFORMATION TO SUPPORT WATER QUALITY MANAGEMENT

Submitted by

Lindsay Melissa Martin

Department of Chemical and Bioresource Engineering

In partial fulfillment of the requirements

for the Degree of Master of Science

Colorado State University

Fort Collins, CO

Spring 2000

COLORADO STATE UNIVERSITY

March 23, 2000

WE HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER OUR SUPERVISION BY LINDSAY MELISSA MARTIN ENTITLED THE ROLE OF DATA ANALYSIS METHODS SELECTION AND DOCUMENTATION IN PRODUCING COMPARABLE INFORMATION TO SUPPORT WATER QUALITY MANAGEMENT BE ACCEPTED AS FULFILLING IN PART THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE.

Committee on Graduate Work

Advisor

Department Head

ABSTRACT OF THESIS

THE ROLE OF DATA ANALYSIS METHODS SELECTION AND DOCUMENTATION IN PRODUCING COMPARABLE INFORMATION TO SUPPORT WATER QUALITY MANAGEMENT

Water quality monitoring is being used in local, regional, and national scales to measure how water quality variables behave in the natural environment. A common problem, which arises from monitoring, is how to relate information contained in data to the information needed by water resource management for decision-making. This is accomplished through analysis of the monitoring data. However, how the selection of methods with which to analyze the data impacts the quality and comparability of information produced is not well understood.

To help understand the connectivity between data analysis methods selection and the information produced to support management, the following tasks were performed: (1) examined the data analysis methods that are currently being used to analyze water quality monitoring data, as well as the criticisms of using those types of methods; (2) explored how the selection of methods to analyze water quality data can impact the comparability of information used for water quality management purposes, and; (3) developed options by which data analysis methods employed in water quality management can be made more transparent and auditable.

These tasks were accomplished through a literature review of texts, guidance and journals related to water quality. Then, the common analysis methods found were applied to the New Zealand Water Quality River Network data set. The purpose of this was to establish how information changes as analysis methods change, and to determine if the information produced from different analysis methods is comparable.

The results of the literature review and data analysis were then discussed and recommendations made addressing problems with current data analysis procedures, and options through which to begin solving these problems and produce better information for water quality management. It was found that significance testing is the most popular method through which to produce information, yet assumptions and hypotheses are loosely explained and alternatives rarely explored to determine the validity and comparability of the results. Other data analysis methods that might be more appropriate for producing more comparable information were discussed, along with recommendations for further research and cooperative efforts to establish water quality data analysis protocols for producing information for management.

Lindsay Melissa Martin
Department of Chemical and
Bioresource Engineering
Colorado State University
Fort Collins, CO 80523
Spring 2000

ACKNOWLEDGEMENTS

This thesis is a compilation of ideas and advice from numerous people. It is not my work alone, and I would like to give credit and thanks to all those who have been involved. I would like to extend my sincere gratitude to my advisor, Dr. Robert Ward, who helped me to develop the idea for this research, and provided a means through which interest and funding became available. He was a constant source of motivation, new thoughts, and enthusiasm for the subject matter. Special thanks go to the providers of that funding, the U.S. Geological Survey, through the interest of Mr. John Klein, co-chair for the National Water Quality Monitoring Council. I would also like to thank the other members of that council, especially Mr. Chuck Spooner, for their sources of information and generosity in allowing me to present my research. I hope they find the information within to be helpful in their endeavors.

Mr. Graham McBride was an excellent guide through the 'murky waters of statistical analysis', and I appreciate not only his expertise, but also the generosity he and Mr. Graham Bryers gave in providing me with data for my research. The format and excellent maintenance of the data made my analysis efforts very easy and straightforward.

I would also like to thank the faculty and students in my department, who have made my experiences here at Colorado State memorable. No one could ask for better

office mates, and friends. The Graduate School Fellowship allowed me freedom to pursue my interests, which grew with the opportunity of being at this university.

I greatly appreciate the love and support of my family, who have offered nothing less during my pursuit of this degree. My parent's lessons of dedication, thirst for knowledge, and performing a job well done served me well throughout these years that I have been away from home. Finally, my love and gratitude are extended Justin Griffith, whose unconditional support and excitement for life has provided immeasurable hope and motivation.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I. Introduction

The passing of the Federal Water Quality Act of 1965 initiated water quality monitoring programs within state water quality management agencies throughout the United States. Before these monitoring programs could mature, a major change in water quality management occurred with passage of the Federal Water Pollution Control Act Amendments of 1972 (commonly referred to as the Clean Water Act today). While appearing to be an update of existing law, the 1972 Act revolutionized water quality management in the U.S. Management of water quality now required large volumes of information about water quality to support sophisticated decision-making (e.g. status of water quality conditions over large spatial and temporal scales, standards violations, and Total Maximum Daily Loads computations).

Requirements of the Clean Water Act included biannual reports, referred to as 305(b) reports and 303(d) lists, from each of the states on water quality conditions. These reports include determinations of designated use support (i.e. is the quality of the water good enough for the typical 'use' of that water, such as swimming or fishing), and lists of waters that are threatened or impaired due to poor water quality. Today, sound data on water quality are becoming increasingly important as numerous lawsuits are directing renewed nationwide attention to the cleanup of water quality problems through the development of total maximum daily loads (TMDLs) for section 303(d) (GAO, 2000).

In order to evaluate the status of their waters, and comply with 305(b) and 303(d) reporting requirements of the 1972 act, states and other entities have collected water quality data and prepared water quality assessments. However, there is a view that the assessments and reporting of this data have provided little indisputable information about the true quality of our nation's waters (PEER, 1999; GAO, 2000). "All too often, monitoring projects are initiated with a minimum of forethought, and result in a collection of poorly-documented data which are never analyzed, [and if they are] provide little or any feedback to resource managers, and contribute little or nothing to our understanding of the systems being monitored" (MacDonald, 1994).

A classic definition of the word monitor is "to watch, observe, or check, especially for a special purpose" (Webster's New Collegiate Dictionary, 1977). Water quality "monitoring" is more than checking to make sure water quality standards are not violated. Monitoring is the process of seeking information about the behavior of water quality variables in the environment (e.g. average conditions, trends, and extremes) (Ward et al., 1986). "Monitoring is performed in support of water quality management and is universally recognized as indispensable for effective management" (Ward et al., 1986).

A common problem, which arises from monitoring, is how to relate information contained in data to the information needed by management for decision-making. For example, if a legal goal from the Clean Water Act is to restore and maintain the nation's water quality, then what information about water quality variables can be used to inform the public and water managers if water quality has been maintained or improved?

A common answer to this problem is to use statistical data analysis methods to produce information from the water quality data. The field of statistics provides an organized approach to quantify the unavoidable uncertainties about the inferences drawn from water quality data (Ward, 1998). Snedecor and Cochran (1980) define statistics as a field that deals with collecting, analyzing, and drawing conclusions from data, and the statistical nature of water quality monitoring has been increasingly recognized (Ward and Loftis, 1983).

Ideally, analysis methods related to specific information goals should be spelled out in advance of collecting data. A way to ensure that comparable information, over time and space, will be developed from water quality data is to thoroughly understand the statistical nature of a monitoring program during the initial design of the monitoring system (Ward, 1998). Knowledge of which statistical tests are most appropriate to obtain the desired information from the collected data plays a role in determining sampling frequencies. Thus, the statistics of a monitoring program are dealt with in a quantitative and transparent manner, before sampling begins. (Ward, et al., 1986) This order of procedure ensures that the appropriate methods for the desired information will be used, and that others who examine the methods will have confidence in the results. It also ensures that the requirements of the analysis methods (i.e. type, quality and amount of data needed) can be determined and used in the design of the monitoring system.

Whether or not this is done, it is common for management to try to produce water quality information from data that were not generated for specific information needs. Often, data are made available from historic or existing monitoring projects, and so analysis procedures must be chosen after the data are collected. How should data

analysis methods be chosen? Answering this question often raises concerns about the validity of the assumptions that are implicit in most statistical analysis procedures, thus calling into question the appropriateness of the analysis procedures chosen. The ad hoc selection of data analysis methods also hurts the validity of the results and the comparability of the information produced. Another, more common concern, is that if the analysis methods are not determined prior to the collection of data, then the analyst has freedom to choose the methods that will produce the outcome that he or she most desires.

Purpose

The purpose of this thesis is to review the current statistical analysis procedures used by a variety of monitoring entities to produce information, and provide some alternative thinking that will serve to strengthen the connectivity between water quality information and the means used to analyze water quality data.

More specifically, the following chapters will: (1) inventory the data analysis methods that are currently being used to analyze water quality monitoring data, as well as the criticisms of current data analysis methods; (2) explore how the selection of methods to analyze water quality data can impact the comparability (i.e. similarity or suitability for comparison) of information used for water quality management purposes, and; (3) offer options by which data analysis methods employed in water quality management can be made more transparent and auditable (i.e. the methods can be reviewed, easily understood, and verified).

These tasks will be accomplished through a literature review of texts, guidance and journals related to water quality monitoring. Then, the common analysis methods found will be applied to a New Zealand Water Quality River Network data set. The purpose being to establish how information changes as analysis methods change, and to determine if the information produced from different analysis methods is comparable. The results of the literature review and data analysis will then be discussed and recommendations made addressing problems with current data analysis procedures, and options through which to begin solving these problems and produce comparable information for water quality management.

Scope

Data analysis, from a water quality management perspective, can be approached from one of two directions: (1) production of information from transparent and auditable data analysis protocols that are comparable over time and space; or (2) exploration of an existing data set to see "what the data say" about water quality conditions in a water body. Statistics are used in both situations, but in different ways. This study addresses the first approach, but realizes that the use of statistics in water quality management often mixes the two.

An argument that often falls out of the above confusion is that there should never be "recommendations" of analysis methods, as this censors the methods that might be used for exploratory data analysis. However, the analysis methods discussed in this thesis will be limited to those methods that are used by water quality management to assess water quality: (1) temporal trends, (2) differences in populations (e.g.

upstream/downstream differences and step trends), and (3) standards violations. These are the three types of information that are most often studied in water quality assessments (Ward, et al., 1990), and which can be used to interpret the quality of the water for regulatory, economic and legal purposes. Therefore, statistics used in modeling analyses (including multivariate analyses, time-series analyses and multiple regression techniques) were not included in this research, as these are used more often as predictive tools.

Chapter II. Criticisms of Water Quality Data Analysis Methods

Water quality assessments are the primary means through which information about our nation's waters is developed. The methods through which data in the medical and behavioral sciences are interpreted are increasingly under fire (i.e. Berger and Berry, 1988; Carver, 1978; Chow and Liu, 1992; Fleiss, 1987; Goodman, 1993; Nunnally, 1960) and some of these criticisms are infiltrating the water quality field. The literature review for this research includes the prevalent criticisms of water quality data analysis.

A recent report by an anonymous group of EPA and other agency employees criticizes the water quality assessments made by states. It states that "inconsistencies in the amounts of waters monitored or evaluated as well as variations in how impairment and designated use attainment are measured, produce a hodgepodge of information that is of little value in determining national water quality trends or comparing water quality among individual States" (PEER, 1999).

Another report produced by the U.S. General Accounting Office reaches similar conclusions about the validity the EPA's *National Water Quality Inventory*, a compilation of all state water quality assessments (305(b) and 303(d) reports). GAO (2000) states that this report can not meaningfully compare information across states because of considerable variation in: (1) the way states select their monitoring sites; (2) the kinds of tests states perform and how the results of these tests are interpreted; (3) the methods used to determine causes and sources of pollution; and (4) the analytical

methods chosen to evaluate water quality (i.e. chemical, physical, or biological properties of water). "By aggregating these states' data, EPA is implicitly suggesting that these data can, in fact, be compared and in doing so is increasing the likelihood that the data will be misused or misinterpreted" (GAO, 2000).

While 15 recommendations are made in the PEER (1999) report to improve the 305(b) reports produced by states, as well as several recommendations by GAO (2000) to improve the usefulness of the *National Water Quality Inventory*, no recommendation is made in either report about how to improve the quality of information produced from states' monitoring systems. One key to this improvement lies in the analysis methods used to interpret the monitoring data. Though analysis methods are rarely questioned, there are a small number of researchers and academics who are questioning the methods used to produce water quality information. This review compiles the arguments brought forth by these critiques.

Similar to PEER (1999), a report of the Virginia Water Quality Academic Advisory Committee (Shabman et al., 1998) makes 17 recommendations to the Virginia Department of Environmental Quality to meet the General Assembly's Water Quality Monitoring, Information and Restoration Act requirements. These recommendations basically cover the water quality assessments used for 305(b) and 303(d) reporting. Several of the recommendations directly address the statistical analysis methods used to produce information from water quality monitoring systems.

Shabman et al. (1998) recognizes the importance of identifying and summarizing water quality trends. At present, the Virginia Water Resources Research Center (VWRRC) is coordinating a research project in which researchers are using improved

(not explained) statistical procedures to perform trend analysis on a watershed scale. Both trend direction and magnitude are stressed, but it is admitted in the report that long-term protocols for statistical analysis and data collection need to be developed. Although it never recommends specific analysis methods for trend detection, in general the report recommends "improved explanations of current use of statistical inference procedures", as well as incorporating the relationship to flow in the analyses for trends. (Shabman et al., 1998)

Currently, Virginia uses EPA's definition of impaired waters, which is defined as an annual violation rate of greater than 10% for numeric water quality standards (referred to as the percentage method). The Virginia Department of Environmental Quality wants to use a binomial procedure to determine probability of violations, due to their small sample sizes, but this is frowned upon by the Virginia Joint Legislative Audit and Review Committee (JLARC), which prefers a standard percent calculation. The committee states that the percentage method is more prone to drawing a false positive inference that a stream segment is impaired, especially if few samples are taken. Use of a binomial distribution is more statistically appropriate, and decreases the chance of a false positive (Type I error). The binomial procedure does not take the actual value or magnitude into account, but if that is of concern, alternatives are suggested. (Shabman et al., 1998)

Santillo et al. (1998) also criticizes the statistics used to determine standards and standards violations. In marine water quality standards, impacts are often based on simple single-species toxicity tests. Essentially what is being determined using such tests is the distribution of tolerance of a given test species. The problem lies in genetic susceptibility of different groups within the same species, as well as the fact that many

responses of species to certain contaminants is not monotonic. Therefore, the criteria for standards violations are not based on the best information.

A third recommendation from the Virginia committee is that the statistical power (i.e. sensitivity) of various temporal sampling patterns should be carefully reviewed in order to design a monitoring program which will optimize analysis opportunities (Shabman et al., 1998). This is a common theme in statistics, and more criticisms of testing without considering power will be discussed below.

The process through which water quality information is produced has become more targeted in the academic field in recent years. Many researchers are criticizing the appropriateness of the actual statistical procedures used to produce the information. From discrediting specific methods for inappropriate use, to rejecting entire categories of methods for inappropriate theory, the typical standard data analysis methods are increasingly being examined in an effort to improve information produced from monitoring.

One critique of incorrect use of methods was prompted by the EPA Guidance for Statistical Analysis of Groundwater Monitoring at RCRA (Resource Conservation and Recovery Act) sites (1989, 1992). In this guidance it is recommended that for a data set with large numbers of nondetects, Poisson prediction limits and Poisson tolerance limits be used. Loftis, Iyer and Baker (1999) prove that neither the Poisson distribution nor associated tolerance or prediction limits should be used with concentration data. "A basic criterion that any model must meet is that it be independent of the system of units, and the Poisson model does not meet that criterion". The problem lies in the fact that the Poisson model does not scale appropriately with changing units, which results from

improper selection of the rate parameter λ in the guidance document. (Loftis, Iyer and Baker, 1999)

Another type of criticism is the issue of statistical power in monitoring design. "Many have noted the lack of attention paid to statistical power in research and monitoring programs" (Santillo et al., 1998). Statistical power is defined as the probability of detecting an effect where one exists, or the sensitivity of the analysis and sampling design to changes in the data. Lack of attention to power has led to experimental designs that seek to minimize the probability of incorrectly identifying an effect when none exists, known as a Type I error (often denoted as α), so as to avoid regulatory regimes that are unnecessarily strict. (Santillo et al., 1998)

However, efforts to minimize Type I errors can lead to increases in Type II errors (denoted as β), an error of accepting the null hypothesis when it is actually false (i.e. not identifying real impacts). "A Type II error could lead to inadequate legislative protection and failure to prevent adverse impacts on the environment or human health. Experiments that fail to identify an effect may lead to acceptance of the null hypothesis (no effect), when the experimental design would have lacked sufficient statistical power to have identified an effect in the first place." (Santillo et al., 1998)

This lack of attention to power considerations draws doubts to the capability of many monitoring programs to detect trends, because too few data points are available to give the analysis much power. "However, designing a monitoring program with enough data points for a decision (say, over 20 years) may result in an environmental impact that is unacceptable". The power also depends on the effect determined to be significant. If the researcher is unable to understand and quantify the extent of impacts caused by

contamination, let alone identify which adverse effects to examine, then a reduction of Type II errors will not reduce scientific rigor of the experiment. (Santillo et al., 1998) On the flip side of this argument is the fact that as databases may grow, tests become too powerful, detecting ever-smaller differences, leading to unimportant differences turning out to be statistically significant (McBride, 1999a).

Another publication, made available on the Internet by the Northern Prairie Wildlife Research Center and the USGS, takes the opposite view and is critical of power analysis. Power analysis, as mentioned above, can be used to determine the sample size needed to have a specified probability (power) of declaring as significant a particular difference or effect (Johnson, 1999). However, when power is determined after a test has been performed to guard against wrongly declaring the null hypothesis to be true, the results can be misleading. This retrospective power analysis, estimated with the actual data used and the observed effect size, is meaningless, as a high p-value will result in a low estimated power (Johnson, 1999). Power analysis programs, however, assume the input values for effect and variance are known, rather than estimated, so they give misleadingly high estimates of power, "as well as requiring three arbitrary parameters, alpha, beta, and effect size". The author states that the questions about the likely size of true effects can be better addressed with confidence intervals than retrospective power analysis. (Johnson, 1999)

The criticism with potentially the most far-reaching impact implies that significance testing is inappropriate for environmental data. Significance testing is the category of statistical analyses that tests a null hypothesis against its alternative, and determines if the outcome is significant evidence against the null or not. "Unfortunately,

when applied in a cookbook fashion, such significance tests do not extract the maximum amount of information available from the data" (McBride, Loftis and Adkins, 1993).

McBride, Loftis and Adkins (1993) claim that significance testing has three problems, which are applicable in environmental monitoring:

1. A conclusion that there is a significant result can often be reached merely by collecting enough samples (increasing sample size increases chance of rejecting the null);
2. A statistically significant result is not necessarily practically significant; and
3. Reports of the presence or absence of significant differences for multiple tests are not comparable unless identical sample sizes are used.

For the past several years, the use of significance testing in the medical profession has been questioned. The argument has been made that the use of arbitrary (i.e. $p < 0.05$) "significance" values does not objectively prove that the data are displaying a characteristic that is not merely chance. In fact, it has been suggested by certain statisticians that p-values are "startlingly prone" to attribute significance to fluke results (Matthews, 1998). Discussions have been raised over the "value" of a p-value, and what it really means in terms of proving anything. Those with less knowledge of statistical theory mistakenly confuse it with the Type I error of hypothesis testing (α), and this link between the two has become standard, but misleading practice (Goodman, 1993). Some data analysts are now questioning the appropriateness of using p-values at all with hypothesis testing (i.e. Goodman, 1993; Berger and Berry, 1988; Matthews, 1998).

The water quality and biology fields are also addressing the confusion over using p-values to support significant findings. Johnson (1999) states that: (1) the p-value is

often used as the probability that the results obtained were due to chance, (2) 1-p is often used as the "reliability" of the result, and (3) p is the probability that the null hypothesis is true.

"Unfortunately, all of these conclusions are wrong. The p-value is the probability of the observed data or more extreme data, given that the null hypothesis is true, the assumed model is correct, and the sampling done randomly" (Johnson, 1999). Determining which outcomes of an experiment or survey are more extreme than the observed one, so a p-value can be calculated, requires knowledge of the intentions of the investigator (i.e. the stopping rule) (Berger and Berry, 1988). "Hence, p, the outcome of a statistical hypothesis test, depends on results that were not obtained, that is, something that did not happen, and what the intentions of the investigator were" (Johnson, 1999). Such information and intentions are often not easily obtained.

Another common mistake in hypothesis testing is that null hypotheses cannot be proved, they can only be rejected. Failing to reject a null hypothesis does not prove that it is true (Johnson, 1999). Especially with small samples, one must be careful not to accept the null hypothesis, as this is a reflection of the lack of power (Johnson, 1999). Even more arbitrary is the designation that a result is "significant" if the p-value falls below some cut-off value, usually given as the acceptable Type I error, α. This means that for an α of 0.05, then a p-value of 0.049 is significant for a one-sided test, whereas a p-value of 0.051 is not (Johnson, 1999). Such a minor difference can be deceptive, as it is derived from tests whose assumptions are often only approximately met (Preece, 1990).

P-values are calculated under the assumption that the null hypothesis is true. Most null hypotheses tested, however, state that some parameter equals zero, or that some set of parameters are all equal. These hypotheses, called point null hypotheses, are almost invariably known to be false before any data are collected (Berkson, 1938; Savage, 1957; Johnson, 1995). If these hypotheses are not rejected, it is usually because sample size is too small (Nunnally, 1960) and power is too low. (Johnson, 1999)

In the field of drug testing, it has been agreed that testing a null hypothesis between means/medians (which is standard practice in water quality data analysis) is not appropriate, as it is evident that the probability of rejecting the null hypothesis increases with sample size (Chow and Liu, 1992). This is due to the fact that the p-value grows smaller as sample size increases. A solution to this problem was given by Good (1982), who proposed that p-values be standardized to a sample size of 100, by replacing the p-value with p*squareroot(n/10), where n is the sample size.

An even more pertinent question would be: why test a null hypothesis at all, if it seems virtually impossible for two different drugs to have the same effect? (McBride, 1998) It has become common practice in drug testing to test whether or not a difference between means/medians might be within a prescribed interval, instead of exactly zero (Chow and Liu, 1992).

Water quality guidance documents, such as the EPA's for statistical analysis of monitoring data at RCRA sites (1989, 1992) often recommend significance testing, such as ANOVA. This type of test can be stated as the following: For the time period given, are the means of a water quality variable equal in all the wells sampled? Or is one or more different from the others? McBride, Loftis and Adkins (1993) point out that as in

drug-testing, we know in advance there will be differences, so why perform the test at all? If there exists a statistically significant difference, this may not translate to a practical significant difference from a management point of view unless power is considered (not the norm).

McBride (1999a) explores this option further. He states that a recurring issue in statistical analysis has been the failure of to use power analysis to select an appropriate sample size so as to minimize the risk either of failing to detect important differences or of detecting the unimportant. "Advocates of power analysis have been increasing in environmental science and management. However, there is discomfort with tests becoming too powerful, i.e. as sample size increases, tests of point hypotheses will tend to detect ever-smaller differences. One response is to de-emphasize the role of tests and rely on confidence intervals." However, McBride (1999a) chooses to support interval testing as a solution to the inappropriateness of testing a point null hypothesis.

Such problems, as discussed above, have led to a "significant test controversy" in the social and behavioral sciences, as well as water quality and biology, with the following remedial measures proposed:

1. Abandonment of testing hypotheses about differences in favor of estimation of differences (Oakes, 1986);
2. Use of interval tests (McBride, 1999a); and
3. Using a combination of estimation and testing with greater emphasis on statistical power in the design of monitoring systems and interpretation of significant test results (Millard, 1987).

McBride, Loftis and Adkins (1993) suggest that the entrenchment of hypothesis testing in the environmental field makes its abandonment unrealistic, but does make several other recommendations related to those in the social and behavioral sciences. One recommendation supports the emphasis on statistical power, stating that both types of errors (Type I and Type II) should be considered when designing a sampling program. "In this way one can seek to have a higher probability of detecting a difference of practical significance (because Type II error is related to the difference in means), corresponding to a particular effect size (chosen by the analyzer), as well as a low probability of raising false alarms". (McBride, Loftis and Adkins, 1993)

Another recommendation is to rely more on interval estimation rather than hypothesis testing. "In trend detection, more information is conveyed by plotting a trend line with confidence limits through a time series than by simplistic yes/no of significance testing." (McBride, Loftis and Adkins, 1993)

The final recommendation by McBride, Loftis and Adkins (1993) refers to interval testing, in which the analysts test whether or not the difference in means is greater than some prescribed interval. "An advantage of this test is that the analyst must state the difference of practical significance to management, also the failure to reject the null no longer induces complacency". This is because the results now mean something, ecologically and environmentally.

Conclusions

The criticisms of data analysis methods in the medical, biology and water quality fields have focused on several key issues. Most of these issues center on the

appropriateness of using hypothesis testing to determine significant results from data. Johnson (1999) even goes so far as to say that "statistical hypothesis tests add very little value to the products of research. Indeed, they frequently confuse the interpretation of data". The arbitrariness and confusion over the meaning of p-values, lack of attention to power, and inappropriate conclusions that the null hypothesis is true all contribute to the ineffectiveness of significance testing. Loftus (1991) "found it difficult to imagine a less insightful way to translate data into conclusions".

Nevertheless, significance testing is still widely used and accepted to develop information from all sorts of data, especially in the water quality field. This prevalence will be demonstrated in the next literature review section. Despite its drawbacks, some advocate more appropriate types of hypothesis testing (i.e. McBride, Loftis and Adkins' (1993) discussion of interval testing), as well as greater attention to the details of the test, including power analysis, sample size and stating the hypothesis. All of these discussions and criticisms help to illustrate the need for more careful attention paid to the selection of analysis methods when the ultimate goal is defensible and comparable information.

Chapter III. Current Water Quality Data Analysis Procedures

The purpose of this literature review is to examine current practice and "state-of-the-art" procedures used to analyze water quality data for information purposes. The review focuses on the use of statistics in literature to produce information, not summary statistics. This information, as discussed in the introduction and scope section, is limited to temporal trends, differences in populations, and standards compliance. The extent of the review covers the major entities involved in water quality monitoring assessments, including the USGS, EPA, private groups and academia, and determines if there are established "standards" of monitoring data, as a whole or within organizational structures. The review covers environmental statistics textbooks, agency publications, water quality reports from state environmental agencies, and the following journals: Journal of American Water Resources Association, Environmental Monitoring and Assessment, Environmental Management, Water Resources Research and Marine Pollution Bulletin.

When beginning this literature review it was thought that there may be "de facto" standards for data analysis developing in the water quality field. Use of the term "standard" is not meant to imply that there is an established set of statistical analysis methods that have been reviewed and recommended for all water quality monitoring situations. However, a large part of this thesis will attempt to establish that there are certain methods that are used time and time again by a variety of monitoring entities, depending on the type of information sought. Conclusions at the end of this literature

review will address whether or not "de facto" data analysis standards are emerging in the analysis of water quality data.

Textbook Guidance for Statistical Procedures to Interpret Water Quality Data

The following tables (Tables III.1 – III.3) summarize information found in three textbooks that are commonly used in the water quality field to determine statistical procedures for data analysis. The purpose here is not to explain the statistical procedures outlined in the text, but to determine which methods seem to be recommended by their inclusion in the text.

Table III.1: *Statistical Methods for Environmental Pollution Monitoring* (Gilbert, 1987)

Information Requirement	Graphical	Parametric Statistics	Nonparametric Statistics
Trends	Time series; CUSUM charts	Regression of deseasonalized data against time with a t-test of hypothesis: slope = 0; Intervention Analysis and Box-Jenkins Model (Autoregressive integrated moving-average time series model)	Mann-Kendall test; Sen's Estimator of Slope; Seasonal Kendall Test/Slope Estimator; Van Belle and Hughes (1984) chi-square test for homogeneity of trend in different seasons; Sen's aligned rank test for trend; test for global trend
Differences in Population		Paired Data: t-test	Paired Data: sign test; Wilcoxon signed rank test; Friedman's test Independent Data: Wilcoxon's rank sum/Mann-Whitney test for two populations; Kruskal-Wallis test for >2 populations
Standards Compliance		Estimating quantiles, proportions, and confidence limits on mean	Estimating quantiles, proportions, and confidence limits on median

Table III.2: *Design of Water Quality Monitoring Systems* (Ward et al., 1990)

Information Requirement	Graphical	Parametric Statistics	Nonparametric Statistics
Trends	Annual Box-and-Whisker plots; time series	Linear Regression, t-test for significance of slope (Snedecor and Cochran, 1980)	Seasonal Kendall Test/Slope Estimator (Gilbert, 1987)
Difference in Populations	Box-and-Whisker; Time series	Student's t-test; Paired t-test; ANOVA (Snedecor and Cochran, 1980); Sample mean or geometric mean with confidence limits (Gilbert, 1987); sample standard deviation with confidence limits (Sachs, 1984)	Seasonal Hodges-Lehman estimator (Hirsch, 1988); Mann-Whitney test, Wilcoxon Signed rank test; Kruskal-Wallis test (Conovor, 1980); Sample median with confidence limits (Gilbert, 1987)
Standards Compliance	Time series plot with Excursion limit	Proportion of Excursions (Ward et al. 1988); Confidence limit on proportions (Gilbert, 1987); Test for equality of proportions (Snedecor and Cochran, 1980); Tolerance Intervals (Conovor, 1980)	Proportion of Excursions (Ward et al. 1988); Confidence limit on proportions (Gilbert, 1987)

Table III.3: *Statistical Methods in Water Resources* (Helsel and Hirsch, 1992)

Information Requirement	Graphical	Parametric Statistics	Nonparametric Statistics
Trends		Regression of Y on T	Mann-Kendall test; Seasonal Kendall test
Difference in Populations	Side-by-Side boxplots; Q-Q plots, Scatterplots with x=y line	Paired data: t-test Independent data: t-test for 2 groups, ANOVA for >2 groups, multifactor ANOVA, two-factor ANOVA Estimating magnitude: confidence interval for difference between means, mean difference t-test; multiple comparison tests	Paired data: sign test; signed-rank test Independent data: Rank-sum test, Kruskal-Wallis test for one-factor >2 groups; ANOVA on ranks; mulifactor test, Blocking – Friedman's test; median aligned-ranks ANOVA; Estimating magnitude: Hodges-Lehmann estimator; median difference sign test
Standards Compliance		Confidence intervals for mean; prediction intervals; confidence intervals for quartiles	Confidence intervals for median; prediction intervals; confidence intervals for quartiles

Recommended Guidance for Statistical Analysis of Water Quality Data

In the search for guidance (i.e. widely applicable and accepted instructions or protocols) on data analysis methods, it appears that no major entity has established a set of comprehensive standards for data analysis procedures. However, the nation's major earth science and environmental agencies, the United States Geological Survey (USGS) and U.S. Environmental Protection Agency (EPA) respectively, have many publications that often serve to guide those who are performing water quality data analysis.

The USGS has no published defined guidance for analysis of water quality data, but does have the largest collection of published water-quality assessments. In these studies, authors often site USGS researcher's publications in their data analysis. For example, Helsel and Hirsch (1992), the textbook mentioned above, is commonly cited as a reference for using the Seasonal Kendall test for detecting trend. In Hirsch (1988), the Hodges-Lehmann class of estimators is found to be robust in comparison to other nonparametric and moment based estimators for determining the magnitude of changes of various constituents between two time periods (step trend). A seasonal Hodges-Lehmann estimator was also developed in this study. By the fact that they are commonly cited in many USGS water quality studies, these types of publications serve as guidance for water quality data analysis in the USGS.

In an academic study, Montgomery and Reckhow (1984) recommend certain techniques for detecting trends in lake water quality, and go on to recommend these procedures for other water bodies as well. This paper stresses the need to formulate a hypothesis, stating that it is only the hypothesis formulated that is being tested. Hence, if information is going to be used in planning and management, one must make sure that the

hypothesis test conducted actually addresses the issue of concern. The authors also recommend plotting the data before choosing the statistical test, as these plots (time series, cumulative sum, histogram, normal probability) can give a visual impression of degree of trend, periodicity, and distribution assumptions. The following statistical tests were recommended according to their data characteristics: (1) for normal and independent data, use two-tailed t or F tests, (2) for normal and dependent data, use a t-test based on the effective number of independent samples, (3) for nonparametric and independent data, use the Mann-Whitney test for a step trend, and Spearman Rho for a linear trend, and (4) for nonparametric, dependent data use correction values on the tests in (3). It was suggested to use statistical tests for dependent data, almost exclusively when dealing with lakes.

Another academic study explores the applicability of the t-test for detecting trends in water quality variables. Montgomery and Loftis (1987) reviewed the effects of non-normality, unequal variances, serial dependence, and seasonality on the performance of the two-sample t-test. The results of this study "suggest that the t-test is robust for non-normal distributions if the distributions have the same shape and sample sizes are equal". It is also robust for unequal variances if the sample sizes are equal. If either of these considerations is not met, as well as the presence of serial dependence or seasonality, then the t-test is not a robust test to detect a step trend.

Another non-agency study, Harcum et al. (1992), recommends using the Seasonal Kendall-tau (SKT) test on monthly data for short periods (less than 10 years) when no serial correlation exists and there is less than 50% missing values. When serial correlation exists, the recommendation is to collapse the data to quarterly values. "Use

the Mann-Kendall test on monthly data with larger records and less than 50% missing values, and collapse to quarterly if greater than 50% missing values." For collapsing, it recommends using median values, and for serially correlated data with long records, a corrected Seasonal Kendall Tau test.

A type of graphical display that is becoming more widely recommended and used in data analysis is the box plot. McGill et al. (1978) describes three variants of the box plot display, which are used in exploratory data analysis and visual summaries. This type of data manipulation does not involve statistics, but an interesting comment from the authors states that "if the notches about two medians do not overlap in this display, the medians are, roughly, significantly different at about a 95% confidence level". Although the authors explain that the user's personal preference is the best criterion for interpretation, this article suggests that graphical displays of data "provide insight into the meaning of the data without the possibility of misinterpretation due to unwarranted assumptions".

Using a study conducted in New Zealand to determine effects of alluvial gold mining operations on benthic invertebrate communities, McBride (1998) demonstrated that traditional point hypothesis tests may not provide satisfactory answers to questions of environmental impact, because they might not be asking or addressing the right questions. Using a standard point hypothesis test, a researcher would examine the null hypothesis that there is no difference at all between the populations being compared, in this case, benthic invertebrate taxonomic richness upstream and downstream from the mining site. The hypothesis is tested by calculating the probability of getting results *at least as different* as those measured merely by chance if this hypothesis were true

(McBride, 1998). If the probability is small (say, less than 5%, $p <= 0.05$), then the null hypothesis is rejected and a "statistically significant" difference has been detected. Using the standard t-test analysis procedure, McBride (1998) found that 4 of the 6 streams showed a "statistically significant" difference between upstream and downstream sites from the mining operation. (See Appendix A for data and results)

The problem with these results lies in the fact that the point null hypothesis says that two of the streams show *no numerical difference at all,* which can hardly be expected in an ecological situation, teeming with natural variability. (McBride, 1998) A better question would be whether there is an "ecological difference" between sites (McBride, 1998). Using the theories of interval testing, an ecological interval could be established corresponding to differences that ecologists deem to be "ecologically significant". If the true difference lies within the interval, the sites would still be "equivalent", and if not, then the sites would be "inequivalent" (McBride, 1998).

It is also possible to set-up the data analysis in two different ways, one with a hypothesis that the differences between populations are equivalent, or one in which they are not (McBride, 1998). When testing using the hypothesis that the sites are equivalent, then only 2 of the 6 streams are found to be inequivalent, or impacted by mining. However, when the hypothesis is that the sites are inequivalent (the difference in means lies outside of the equivalence interval), only one of the streams is deemed equivalent, therefore mining has impacted 5 streams. (McBride, 1998) The information produced is very different, and reflects an emphasis or non-emphasis on environmental protection, a key point to environmental management. Testing the null hypothesis that the streams are equivalent protects the environmental user's risk, resting the "burden of proof" on the

monitoring system to show that an impact has occurred. However, the latter approach of testing a null hypothesis of inequivalence is a more 'precautionary' approach, assuming the stream has been impacted, unless proven otherwise (McBride, 1998). These results show the importance of complete understanding of the implications behind each hypothesis to management decision making. These results also show the importance of determining the test hypothesis before analysis, as information can change depending on the structure of the hypothesis.

McBride (1998) also explores the differences between hypothesis testing and using Bayesian statistics, which establishes a degree of belief in the hypothesis, then updates the belief in light of the data. "Using a Bayesian test procedure only depends on the data obtained, and can be viewed as a weight for or against equivalence, which might be the most direct answer for the original question asked: are upstream and downstream sites of the mining operation equivalent?" (McBride, 1998)

The largest collection of guidance for data analysis was found in publications by the U.S. Environmental Protection Agency. Guidance has been published by the EPA for the states' submittal of 305(b) and 303(d) reports. However, no specific statistical methods appear to be endorsed by the organization for these reports. For 303(d), it is stated that states should determine threatened waters by data showing a statistically significant declining trend. The state's report should describe how the trend was determined, but no particular requirements for trend detection are mentioned (EPA, 1998). The EPA does state that it prefers to base listing decisions on monitored data for all their waterbodies, though it recognizes most states do not have a comprehensive enough monitoring program. This recommendation is due to the EPA's desire that listing

decisions be based on sound, high-quality, scientific information. These 303(d) listing seem to be closely linked and dependent on the states' 305(b) reports and designated use support determinations.

The Guidelines for Preparation of State Water Quality Assessments (305(b) Reports) and Electronic Updates for the 1998/2000 Reporting Cycle (EPA, 1997a) advises entities to document summary statistics for use support and the approaches used to identify causes and sources of impairment (i.e. standards violations), along with confidence levels. The major reporting format is miles and acres of designated threatened, use supporting or non-supporting water bodies, but no mention is made of how statistically sound inferences, from limited samples, are to be applied to an entire water body. However, the EPA does make recommendations on how states should determine use support numerically and narratively. (See Tables III.4 and III.5)

This literature review found that the EPA mainly publishes guidance that helps the states and other reporting entities compile and interpret information to support EPA rules and programs. One such guidance is the *Information Collection Rule: Draft Data Analysis Plan* (EPA, 1997b). One objective of the ICR is to collect data on specific water quality constituents (i.e. DBP precursors, disinfectants) and use this data to characterize source water parameters that influence disinfection byproduct (DBP) formation, refine models for predicting DBP formation, and establish cost-effective monitoring techniques. The guidance discusses the need to characterize baseline conditions and predict changes and impacts, but does so by asking questions about how constituent data is to be evaluated without providing answers. The only statistics mentioned in the document are options for summary statistics used to characterize the data, i.e. averages, ranges and

percentiles with confidence intervals, and cumulative probabilities. This guidance does suggest that the decision on what statistical approach to use for aggregating the occurrence data will depend on the specific question which is trying to be addressed.

Table III.4: Recommendations from EPA's 305(b) Guidance for Interpreting Water Quality Criteria (EPA, 1997a)

Parameter type	Fully Supporting	Impaired	
		Partially Supporting*	Not Supporting*
Conventional	Criterion exceeded in ≤ 10 percent of measurements	Criterion exceeded in 11 to 25 percent of measurements	Criterion exceeded in > 25 percent of measurements
Toxicants	No more than 1 exceedance of acute and chronic criterion within 3 yr. period (at least 10 measurements over 3 yr.)	More than one exceedance of acute or chronic criterion, but in ≤ 10 percent of the samples	Acute or chronic criterion exceeded in > 10 percent of the samples
Biological integrity	Reliable data indicate functioning, sustainable biological assemblages none of which have been modified significantly beyond the natural range of the reference condition	At least one assemblage indicates moderate modification of the biological community compared to the reference condition.	At least one assemblage indicates nonsupport. Data clearly indicate severe modification of the biological community compared to the reference condition.
Habitat	Reliable data indicate natural channel morphology, substrate composition, bank/riparian structure, and flow regime of region. Riparian vegetation of natural types and of relatively full standing crop biomass.	Modification of habitat slight to moderate usually due to road crossings, limited riparian zones because of encroaching land use patterns, and some watershed erosion. Channel modification slight to moderate.	Moderate to severe habitat alteration by channelization and dredging activities, removal of riparian vegetation, bank failure, heavy watershed erosion or alteration of flow regime.
Toxicity - aquatic or sediment	No toxicity noted in either acute or chronic tests compared to controls or reference conditions	No toxicity noted in acute tests, but may be present in chronic tests in either slight amounts and/or infrequently within an annual cycle.	Toxicity noted in many tests and occurs frequently.
Bacteria	E.coli and enterococci - Geometric mean of samples taken should not be exceeded and single sample does not exceed the maximum allowable density Fecal coliform - geometric mean does not exceed 200 per 100ml based on at least five samples in 30 day period and not more than 10 percent of the total samples taken during any 30 day period have a density that exceeds 400 per 100ml	E.coli and enterococci - geometric mean met; single sample criterion exceeded during the recreational season Fecal coliform - geometric mean met; more than 10 percent of samples exceed 400 per 100ml	Geometric mean not met

Table III.5: Recommendations from the 305(b) Guidance for Making Use Support Determinations (EPA, 1997a)

Designated Use	ATTAINING		IMPAIRED	
	Fully Supporting	**Fully Supporting, but Threatened**	**Partially Supporting**	**Not Supporting**
Aquatic Life	No impairment indicated by all (available) data types	No impairment indicated by all (available) data types, but: - one or more categories indicate an apparent decline in ecological quality over time - potential water quality problems requiring additional data or verification - other information suggests a threatened determination	Impairment indicated by one or more data types and no impairment indicated by others	Impairment indicated by all data types
Primary Contact Recreation Use	Bathing area closure: - No bathing area closures or restrictions in effect during reporting period Bacteria: - See table 1		Bathing area closure: - On average, one bathing area closure per year of less than 1 week's duration Bacteria: - See table 1	Bathing area closure: - On average, one bathing area closure per year of greater than 1 week duration, or more than one bathing area closure per year Bacteria: - See table 1
Drinking Water	Contaminants do not exceed water quality criteria and/or drinking water use restrictions not in effect	Contaminants are detected, but do not exceed water quality criteria and/or some drinking water use restrictions have occurred and/or the potential for adverse impacts to source water quality exists	Contaminants exceed water quality criteria intermittently and/or drinking water use restrictions resulted in the need for more than conventional treatment with associated increases in cost.	Criteria exceed water quality criteria consistently and/or drinking water restrictions resulted in closures.
Fish/Shellfish Consumption	No fish/shellfish restrictions or bans are in effect.		"Restricted consumption" of fish in effect or a fish or shellfish ban in effect for a subpopulation that could be at potentially greater risk, for one or more fish/shellfish species.	"No consumption" of fish or shellfish ban in effect for general population for one or more fish/shellfish species or commercial fishing/shellfishing ban in effect

The Monitoring Guidance for the National Estuary Program (EPA, 1992) is another example of guidance to support an EPA program. One of the five steps outlined in the recommended design framework is to "establish testable hypotheses and select statistical methods". The guidance states that "the recommended procedure for ensuring that sufficient information and the right type of information is developed in the monitoring program is to specify, prior to the collection of any samples, the statistical model that will be used to analyze the resulting monitoring data, and to specify testable hypotheses". The selection of the hypothesis is discussed in its relationship to the objective of the monitoring program and the question needing to be answered. As far as a recommendation of statistical methods, the guidance provides a list of textbooks on monitoring design and statistics: *Sampling Design and Statistical Methods for Environmental Biologists* (Green, 1979), *Statistical Methods for Environmental Pollution Monitoring* (Gilbert, 1987), *Sampling Techniques* (Cochran, 1977) and *Statistical Principles in Experimental Design* (1971). It also recommends several general statistics books, and multivariate statistics books. This guidance also gives a good explanation of the theory of statistical power and its importance as an evaluation method for the ability of a monitoring program to detect statistically significant differences.

The most comprehensive of EPA's guidance in terms of statistics is the *Monitoring Guidance for Determining the Effectiveness of Nonpoint Source Controls* (EPA, 1997c). This document dedicates a whole section to statistics, covering estimation and hypothesis testing, characteristics of environmental data, and recommendations for selecting statistical methods. Many of the recommended methods are adapted from the textbook *Design of Water Quality Monitoring Systems* by Ward et al. (1990), which was

reviewed at the beginning of this chapter. This guidance covers the theory and application of summary statistics, graphical data display, evaluation of test assumptions (i.e. tests for normality), and provides a list of references and useful software for data analysis. The guidance also covers regression techniques, analysis of covariance, correlation coefficients, multivariate analysis, and extreme events. The statistical tests covered and explained in the guidance are listed in Table III.6 below.

Table III.6: Recommended Statistical Analysis Tests for Determining Effectiveness of Nonpoint Source Controls (EPA, 1997c)

Tests for One Sample or Paired data	Student's t-test Wilcoxon Signed-Rank test Sign test
Two-sample tests	Two-sample t-test Mann-Whitney (Wilcoxon's rank-sum) test
Magnitude of differences	Confidence interval of differences between means Hodges-Lehmann Estimator
Comparison of >2 Independent Samples	ANOVA (one-factor and two-factor) Kruskal-Wallis test Ranked transformed ANOVA Friedman test
Multiple comparisons	Tukey's method Bonferroni t-test Duncan's multiple range test Gabriel's multiple comparison procedure (REGW) mutiple F-test and range test Scheffe's multiple-comparison procedure Waller-Duncan k-ratio test
Monotonic Trends	Mann-Kendall test Seasonal Kendall test

The EPA has established guidelines for *Statistical Analysis of Groundwater Monitoring Data at RCRA (Resource Conservation Recovery Act) Facilities* (EPA, 1989;1992). Five statistical methods were outlined in the Final Rule: (1) a parametric analysis of variance (ANOVA), (2) a nonparametric ANOVA based on ranks, (3) using tolerance levels or prediction levels from background data and then comparing each constituent to the upper levels, (4) a control chart approach which gives control limits for each constituent, and then comparing sample values to these limits, and (5) another

statistical method submitted by operator and approved by the Regional Administrator. The guidance provides flowcharts to help operators decide which method to use, as well as ways to check distribution assumptions and homogeneity of variance. The 1992 addendum adds several recommendations. It addresses more methods for checking assumptions for statistical procedures and homogeneity of variances, as well as recommendations for handling nondetects. For comparison of populations (wells to background data), the guidance addendum adds (1) the nonparametric Kruskal-Wallis test, and (2) the nonparametric Wilcoxon Rank-Sum (Mann-Whitney) test for two groups.

The EPA also has research publications that can be viewed as endorsements for particular methods. In Loftis et al. (1989), seven statistical tests for trend were evaluated under various conditions and performance was compared using actual significance level and power. The evaluations resulted in the following recommendation by the authors: for annual sampling use the Mann-Kendall test for trend, and for seasonal sampling, use either the Seasonal Kendall test or the Analysis of Covariances (ANOCOV) on ranks test.

A guidance document for determining improvements from agricultural nonpoint source control programs was developed and published by North Carolina State University for the EPA (Spooner et al., 1985). The authors give recommendations on monitoring design, appropriate hypotheses, data requirements, assumptions, and testing procedures. For time trend analysis without correction for meteorological variables, Spooner et al. (1985) recommends the Students t-test, graphical/regression analysis of the concentration versus BMP application level, or the use of a Quantile – Quantile (Q-Q) plot. Time trend analysis corrected for stream flows should use separate linear regressions of

concentrations versus flows for the pre- and post-BMP periods. Then, the slopes can be compared for equality. For upstream/downstream analysis, again the recommendations are to use Students t-test, Q-Q plots, or linear regression of concentrations versus BMP implementation level or flow. Finally, for paired watershed analysis, the authors recommend linear regressions of the concentrations for the treatment versus the control watersheds for both the calibration and land treatment time periods. A Students t-test can be performed to determine if the "predicted treatment watershed values at the mean control watershed concentration decreses over time" (Spooner et al. 1985).

Although not a published document, the EPA is working on a *Technical Guidance on Monitoring and Data Interpretation to Support Implementation of Water Quality Standards* (EPA, 1999). This document is in outline form, but has the objectives of: (1) improving the scientific basis of decisions to characterize waters as being compliant, threatened or impaired; (2) providing guidance for developing an assessment methodology; and (3) promote functional integration of monitoring and data sharing, data analysis and interpretation across state programs responsible for water quality characterization and decision making. The document was written to support state water quality stream standards, 305(b) and 303(d) requirements. Although it makes no statistical recommendations itself, the guidance asks that protocols be established for determining standards compliance and determining trends, as well as makes recommendations for the characteristics of the data needed to support decision making (i.e. coverage, number of samples, gaps in record, frequency of samples).

With the exceptions discussed above, attempts to produce standard sets of guidance procedures for water quality data analysis are relatively few and uncoordinated

between agencies. In the field of groundwater monitoring, Adkins (1992) states that "due to the wide variety of information needs and site conditions, it is impractical to expect a single data analysis protocol to be suitable for all groundwater quality monitoring systems…[and that] no generally acceptable design framework for the development of groundwater quality data analysis protocols exists today". Therefore, instead of recommending specific analysis procedures, Adkins (1992) presents a framework for the development of groundwater quality data analysis protocols.

The next step of the literature review was to determine what the actual current use of statistics is in water quality data analysis. Although general standard methods for water quality monitoring analysis may not be published, it is hypothesized that they are established through common practice, especially within organizations and types of monitoring entities.

Peer Reviewed Water Quality Assessments

This section serves to establish the current use of statistics, beyond guidance, in the water quality field. To gain a comprehensive view of the use of statistics, recent issues of five major environmental refereed journals were examined: Journal of American Water Resources Association, Environmental Monitoring and Assessment, Environmental Management, Water Resources Research and Marine Pollution Bulletin. The peer-reviewed studies included here are limited to those which sought information related to environmental management: Temporal Trends, Differences in Population (including upstream/downstream differences, before/after differences, and spatial differences), and Standards Compliance.

Trend Analyses

Most trend analyses were performed with non-parametric tests for trend, to avoid complications in the data set and assumptions of normality, making the tests more robust. The most popular analysis was the Seasonal Kendall Tau (seasonal extension of the nonparametric Mann-Kendall) test for monotonic trend, used in 12 out of the 19 studies where trend was determined (highlighted in gray, Table III.7). It is especially popular with USGS studies. The USGS is also very thorough about performing the test on both the original data and flow-adjusted concentrations, but only if a strong correlation exists between concentration and flow. All studies reviewed which dealt with trend detection are summarized in Table III.7. A few of the studies used alternative procedures to determine trends, as summarized below.

Lavenstein and Daskalakis (1998): The Kendall-tau nonparametric test for linear correlation was used to determine trends in constituent data from 3 Mussel watch programs.

Stoddard et al. (1998): In order to infer regional trends (over several monitoring sites and data sets), this study employed a variation of meta-analysis. Trends were assessed through the use of the Seasonal Kendall-tau test, and then the resulting statistics were combined through a technique analogous to ANOVA, to produce quasi-regional estimates of changes for key chemical variables. This technique is referred to as Analysis of Chi-Squares.

Table III.7: Water Quality Assessments Involving Trend Detection

Author	Monitoring Entity	Distribution Assumption	Actual Hypothesis Stated	Test Used
Clow and Mast (1999)	USGS	NP	None stated	Season Kendall Tau or Mann-Kendall
Baldys, Ham and Fossum (1995)	USGS	NP	Null hypothesis of no significant trend	FAC Season Kendall Tau or Mann-Kendall
Mattraw, Scheidt and Federico (1987)	USGS, NPS and SFWMD	NP	None stated	FAC Season Kendall Tau or Mann-Kendall
Rinella (1986)	USGS	NP	None stated	FAC Season Kendall Tau or Mann-Kendall
Berndt (1996)	USGS	NP	None stated	Season Kendall Tau or Mann-Kendall
Mueller (1995)	USGS	NP	None stated	FAC Season Kendall Tau or Mann-Kendall
Mueller (1990)	USGS	NP	None stated	FAC Season Kendall Tau or Mann-Kendall
Snyder et al. (1998)	Academia	NP, Parametric	null = no tendency for one sampling location to have nutrients greater than another location	Duncan's new multiple range test (Ott, 1988) - test of the diff. In means of multiple pops., % reduction of means
Stoddard et al. (1998)	EPA, Academia, Vermont DEC	NP	None stated	SKT. Analysis of Chi-squares and meta-analysis
Pinsky et al. (1997)	EPA, Academia	NP, Parametric	None stated	Auto-regressive first order process, comparing means/medians
Takita (1998)	Susquehanna	NP	None needed	Double mass comparison
Havens et al. (1996)	SFWMD	Parametric	None stated	Satterwaite's t-test
Dennehy et al. (1995)	USGS	NP	Null states that no trend exists	LOWESS (to highlight patterns), FAC SKT
Butler (1996)	USGS	NP, Parametric, Parametric, NP	Null means there is no trend or no sig. diff between means/medians	FAC-SKT (periodic & monthly), FAC LR (annual), Step Trend two sample t-tests, Wilcoxon Rank Sum
Smith, Alexander and Wolman (1987)	USGS	NP	None stated	SKT and FAC SKT
Vaill and Butler (1999)	USGS	NP	Null hypothesis of no trend	monotonic trends: SKT and FAC SKT, Sen Slope estimator, Lowess to determine in what part of the record the trend occurred, Step trends: Parametric 2-sample t-test and NP Wilcoxon rank-sum test applied to raw data
Heiskary, Lindbloom and Wilson (1994)	Minnesota Pollution Control Agency	NP	Null hypothesis of no trend	Kendall's tau-b (Gilbert, 1987)
Lavenstein and Daskalakis (1998)	NOAA	NP	None stated	Kendall-tau test for linear correlation
Brown et al. (1999)	NOAA	NP	None Stated	Spearman-rank Correlation method, meta-analysis

Snyder et al. (1998): This study used Duncan's new multiple range test to test the difference in means of multiple populations. Although this seems like a difference in population's study, the % reduction in means was used to support evidence of temporal trends.

Pinsky et al. (1997) and Takita (1998): These two studies didn't use statistical hypothesis tests to determine trends, and instead used information from the actual data. Pinsky et al. (1997) just compared means/medians and inferred trends with an auto-regressive first order process. Takita (1998) used plotting procedures to determine the data's approach towards a trend, called double mass comparison.

Havens et al. (1996): This study from the South Florida Water Management District used Satterwaite's t-test to determine trend.

Butler (1996): In order to determine smaller trends in the data without the assumption of a monotonic trend, this study used a step-trend analysis, using two-sample t-tests, along with the Wilcoxon Rank-Sum test. Butler (1996) did use the flow-adjusted SKT test for periodic and monthly data trends, but for annual data, used a linear regression technique.

Vaill and Butler (1999): Although this study performed the standard Seasonal Kendall test for monotonic trend analysis, it also looked at step trends where a known event occurred at a specific time in the watershed. The author used a parametric two-sample t-

test where the raw data was distributed normally, and the Wilcoxon Rank-Sum test where the data was not normal.

Heiskary, Lindbloom and Wilson (1994): Trends were assessed using Kendall's tau-b statistical test, a non-parametric procedure that computes correlation coefficients between variables. The null hypothesis was stated as no trend, and the strength of the relationship was a function of both the correlation coefficient and the number of years of measurement. This study also determined the sampling frequency needed to maximize the power of detecting a significant change (established as weekly to allow a 70% chance of detecting a 20% change over 10 years). To check the validity of their results, the authors attempted to corroborate the results using trophic status, user perception, watershed and modeling information.

Brown et al. (1999): The Spearman Rank correlation method, a bivariate nonparametric procedure and a meta-analysis procedure (discussed further in Chapter V) were used to examine relationships among chemical concentrations in sediment and fish tissue. "Although the temporal trends in this study do not conform in the strictest sense to meta-analysis assumptions of independence, it was assumed that the compartments analyzed were distinct enough for synthesis into a single test for trend." This was accomplished by taking the significance levels for the Spearman rank correlations, transforming them into z-values, combining them and transforming them back into a single p-value. This resultant significance level gave an indication of consistency across compartments and statistical certainty with which a trend exists for a contaminant at a site.

Differences in Populations

There were a greater variety of tests chosen to determine differences in population. Three major groups of analyses prevailed: (1) using Signed Rank, Rank Sum or variations of those procedures, (2) using cluster type analyses and (3) using ANOVA or variations. The most popular tests were the Wilcoxon Rank-sum/Mann-Whitney test or its extension for more than 2 populations, the Kruskal-Wallis test (8 out of 20 studies reviewed, light gray highlight in Table III.8) and the Analysis of Variance test (ANOVA used in 5 out of 20 studies, dark gray highlight in Table III.8). Most studies tested for normality before choosing a difference test, though some just assumed nonparametric statistics should be used. Almost all the tests used were for nonparametric distributed data. With the exception of Dennehy et al. (1995), no hypotheses were given. But it was evident by the testing that all performed a significance test with a null hypothesis of the means/medians between groups being equal.

The USGS studies seemed to prefer the Wilcoxon Rank-Sum (Berndt, 1996; Abeyta and Roybal, 1996) or Kruskal-Wallis test (Abeyta and Roybal,1996; McMahon and Harned, 1998; Mueller, 1995; Dennehy et al., 1995). All of the studies reviewed are summarized in III.8. Specific explanations of some of the more unique analysis methods are provided below.

Arthur, Coltharp and Brown (1998): This was the only study that used the Wilcoxon Signed-Rank test for differences, as opposed to the common Wilcoxon Rank-Sum test.

Table III.8: Water Quality Assessments Involving Differences in Populations

Author	Monitoring Entity	Distribution Assumption	Actual Hypothesis Stated	Test Used
Younos et al. (1998)	VWRRC, Academia	NP	None stated	Wilcoxon Test (Hollender & Wolfe 73)
Arthur, Coltharp and Brown (1998)	Academia	NP	None stated	Wilcoxon Signed Rank
Berndt (1996)	USGS	NP	None stated	Wilcoxon Rank-Sum
Pinsky et al. (1997)	EPA, Academia	NP, Parametric	None stated	Wilcoxon Rank-Sum, Chi-Square test of hypothesis of equal proportions in population
Abeyta and Roybal (1996)	USGS	NP, NP, NP, Parametric	None stated	Wilcoxon Rank-Sum, Kruskal-Wallis, ANOVA, ANOVA & paired t-tests
Sample et al. (1998)	USDA NRCS	NP, NP, NP	None stated	Rank Sum, Signed Rank, Hodges-Lehmann Estimator
McMahon and Harned (1998)	USGS	NP	None stated	Kruskal-Wallis, and Tukey's Multiple Comparison
Mueller (1995)	USGS	NP	None stated	Kruskal-Wallis
Koebel, Jones and Arrington (1999)	SFWMD	NP, NP	None stated	TSS, Turbidity, Nutrients - Kruskal-Wallis, Dunn's test, ANOVA & paired t-tests
Momen et al. (1997)	Academia	Parametric, Parametric	None stated	Tukey's multiple comparison for mean separation, ANOVA (temporal and spatial)
Takita (1998)	Susquehanna	NP	None needed	Plotted Annual Loads vs. Discharge Ratio
Dennehy et al. (1995)	USGS	NP	Null states that no difference exists	Kruskal-Wallis test
Snyder et al. (1998)	Academia	NP?	None stated	Friedman's test (Gilbert, 1987), Cluster Analysis (Davis, 1986), Cross-Correlation Analysis
Stoe (1998)	Susquehanna	Parametric?	None stated	PCA, Cluster analysis, Habitat Assessment scores and Biological Condition scores
Nimmo et al. (1998)	USGS, EPA, Academia, CDOW	Parametric	None stated	ANOVA & paired t-tests, Student-Newman-Keuls method of separating means
Colman and Clark (1994)	USGS	NP	None stated	ANOVA
Rinella (1986)	USGS	NP	None stated	Tukey's multiple comparison
Kennedy (1995)	TxDOT, North Central Texas COG	NP	None stated	Kruskal-Wallis test, Mann-Whitney test
Kress, Hornung and Herut (1998)	Israel Oceanographic and Limnological Research	Parametric	None stated	GLM least squares, t-test, Mann-Whitney a-parametric test
Brown et al. (1999)	NOAA	NP	None stated	GT2 multiple comparison method

Brown et al. (1999): The relative concentration of contaminants in sediment and fish tissue were compared statistically using the GT2 multiple comparison method, which is equivalent to performing a one-way ANOVA followed by a multiple-range test. In

graphical displays of GT2 comparison intervals, those that do not overlap are significantly different at the ($p<=0.05$) level.

Kress, Hornung and Herut (1998): The purpose of this study was to assess the influence of dumping on the trace metal contents of deep-sea benthos. To compare the populations from the dump sites to the fauna population at a control, the authors used a general linear model of least squares, a t-test, and a Mann-Whitney a-parametric test at the 95% confidence level.

Kennedy (1995): The Texas Department of Transportation used nonparametric procedures to determine differences in stormwater runoff. Their specific purpose was to determine whether a significant difference could be detected among runoff from four different landuse categories. They used the Kruskal-Wallis test to determine if there was a difference among the four sites, and then the Mann-Whitney test for each combination of two-sites to determine the site of greatest difference.

Pinsky et al. (1997): In this study academia and the EPA assumed independence of the wells that were sampled. For analysis of the proportions of wells with a certain characteristic, the standard normal approximation to the binomial distribution was used to generate confidence intervals, and a Chi-Square test was used to test the hypothesis of equal proportions in two populations. The Wilcoxon Rank-Sum was used to test whether the distribution of a quantitative variable was the same in two populations of wells. No tests were performed within a well because of non-independence.

Sample et al. (1998): The only USDA NRCS study reviewed used the general Rank-Sum and Signed-Rank tests, along with a Hodges-Lehmann estimator to determine the magnitude of increasing or decreasing water quality degradation.

Koebel, Jones and Arrington (1999): This study by the South Florida Management district tried to determine water quality impacts from canal backfilling. The analysts used several different tests for to detect differences in populations, including the Kruskal-Wallis test, Dunn's test for post hoc multiple comparisons of site differences, and ANOVA with paired t-tests.

Takita (1998): This study's purpose was to quantify nutrient and sediment transport in the Susquehanna River Basin. To analyze for annual variation in loads, the author did not even use statistics, but instead used a graphical procedure of plotting Annual Loads vs. the Discharge Ratio. If a certain site's plot differed from the baseline plot, then a change in population was assumed to have taken place.

Snyder et al. (1998): To determine the impact of riparian forest buffers on agricultural nonpoint source pollution, the authors used a cluster analysis and cross-correlation analysis to support evidence of differences.

Stoe (1998): This study utilized a cluster analysis called Principal Components Analysis (PCA) for water quality, along with a non-statistical Habitat Assessment Score and

Biological Condition Score for an ecological assessment of differences between sampling sites.

Standards Compliance

Determination of standards compliance was not commonly sought via statistical tests in the research type assessments that were reviewed (see Table III.9 for summary of assessments which involved standards compliance). Therefore, part of this literature review attempted to describe how states generate this information for their 305(b) and 303(d) reporting requirements, especially in light of the current 303(d) listings and TMDL debate. Many states do not publish their assessment methodologies, so personal communication via the phone and/or email was the primary venue through which such information was gathered. The purpose was to try and establish if there are common methods used by the states for their water quality assessments, not to document every detail of their assessment methodology. It was found that documented analysis methods or statistical tests are rarely used to determine use support assessments or standards violations. Often only simple "percentage of standard exceedences" is used to assess a water body, along with subjective evaluation of the waterbody according to narrative criteria.

Table III.9: Water Quality Assessments Involving Standards Compliance

Author	*Monitoring Entity*	*Distribution*	*Hypothesis Stated*	*Test Used*
Berndt (1996)	USGS	NP	None stated	% exceedence of MCL, highest means reported
Lapp et al. (1998)	Academia	NP	None stated	observed mean does not exceed DW standard in Canada
Nimmo et al. (1998)	USGS, EPA, Academia, CDOW	Parametric	None stated	average concentrations compared to chronic 4-day aquatic life criterion (USEPA)
Bexfield and Anderholm (1997)	USGS	?	None stated	compared daily and quartile concentrations to standards

State Determinations of Designated Use Support

New York: Judgements are made on use support according to narrative criteria established by the state. New York stated that "the bulk of Priority Waterbody List (PWL) information is reflective of *evaluation* as opposed to *monitoring* efforts. This report did not qualify how the area of effect (i.e. stream miles) is determined for each segment reported. They are currently implementing a rotating basin approach for future assessments. (NYS Department of Environmental Conservation, 1998)

New Jersey: Judgements on use support are qualified by monitoring data and criteria developed by the state. No statistical tests are used. However, the protocol for determining use support is documented thoroughly. For example: for recreational use support, data collected over 5 years was compared to the NJ Surface Water Quality Standard criteria for fresh water streams, and use support determined according to the criteria listed below in Table III.10.

Table III.10: New Jersey Recreational Use Support Criteria

Use Support	**Assessment Criteria**
Full Support	The fecal coliform geometric avg. was <200 MPN/100ml and <10% of individual samples exceeded 400 MPN/100ml
Partial Support	Fecal coliform geo. Avg. was <200 MPN/100ml but >10% of samples exceeded 400 MPN/100ml
No support	Fecal coliform geo. Avg. >200 MPN/100 ml and >10% of samples exceeded 400 MPN/100ml

New Jersey also established its miles affected according to the criteria that the number of miles is the distance between the 2 monitoring points plus 1000 feet upstream. Other use support designations and trends were reported, but no protocol was documented for their determination. (NJ Department of Environmental Protection, 1998)

Region III (Delaware, Pennsylvania, Maryland, Virginia, West Virginia, District of Columbia): Criteria for use support assessment are those recommended by the EPA for 305(b) reports (see Table III.4 and III.5). Some states use biology to determine use support, following the EPA's Rapid Bioassessment Protocol. "By and large, simple percentages of standard violations are used to make a judgement call for water body assessments" (Barath, 2000).

Oklahoma: This state delineates all of their criteria for use support determination, with most criteria being comparisons of monitored data to standards. For example: Oklahoma uses the EPA recommendations for numerical parameters (full support = <10% violations, partial support = >11% but <25% violations, and no support = >25% violations). At least ten samples are required for this determination in streams, and 20 vertical profiles in lakes. However, fewer can be used if exceedence is assured. Any monitoring site shall not represent more than 10 wadable stream miles, or a lake area more than 250 surface acres. (Oklahoma Water Resources Board, 1999)

Arizona: No trends are evaluated, and no statistical tests are used. The use support criteria (see Appendix B) are enumerated from Arizona DEQ (2000). Arizona also uses macroinvertebrate-based bioassessment criteria to determine use, generally following EPA's guidelines. However, this Index of Biological Integrity (IBI) is not statistically based, it uses a scoring system and percentiles. No water body assessed as partially supporting or non-supporting based solely on biocriteria will be placed on the state's

303(d) list prior to identification and cause of the impairment, as it could be the results of natural phenomenon. (Marsh, 2000)

California: Individual regions do not provide information about how they determine use support. The only known protocol is for Los Angeles, which uses the criteria recommended by the USEPA (see Table III.4 and III.5). (Richard, 2000)

Hawaii: Use support is determined partially by comparing bacteria and chemical water quality data to state standards. For those categories which don't have applicable state standards, narrative criteria were created for judgement decisions instead of numerical/statistical based decisions. (Teruya, 2000)

Virginia: Criteria for use support enumerated is by the state (see Appendix C). The actual numerical/narrative decision protocol follows the EPA recommended criteria for use support determinations. Assessment decisions are based on both monitored and evaluated data. Virginia also sets protocols for determining affected areas, e.g. stating that no station shall represent more than 10 miles of wadable stream. This determination is a judgement-based decision taking several enumerated factors into account. (Virginia Department of Environmental Quality, 1999)

South Carolina: This state uses the EPA's recommended assessment criteria for 305(b) reporting (See Tables III.4 and III.5 above). (Kirkland, 2000)

Florida: As a portion of Florida's efforts, the state has adopted an Environmental Mapping and Assessment Program (EMAP) type of statistical analysis. The goal is to determine the overall conditions of water bodies within a geographical area. For example, the state will make statements such as, "With a confidence level of .90, the median value for NO3 in small lakes in north central Florida is (say) 1.3 mg/l plus or minus 0.4 mg/l. The state has been broken into 20 geographical units based on hydrologic drainage basins. These analyses will be performed for six resources. They are confined ground water, unconfined ground water, small lakes, large lakes, high order streams and low order streams. A sister organization in the state is conducting a similar analysis for Florida's estuaries. (Copeland, 2000)

Tennessee: This state generally follows the EPA's recommendations for use assessments (See Tables III.4 and III.5 above), but has some discretion in the "magnitude and duration" of water quality standard violations. (Denton, 2000)

North Carolina: Use support for 305(b) and 303(d) listing are based on monitored and evaluated data, with more confidence placed on monitored data. Biological indexes and physical/chemical data are used to determine use support, similar to the procedures Arizona uses (See Appendix B). However, biological data/indexes take precedence over chemical/physical data when determining use support. (Swanek, 2000)

Kentucky: Kentucky's approach is a combination of targeted sites and random survey sites. They mainly use biological data to determine use support. Many of their water

quality stations are at sites also sampled biologically. However, there are a few sites, mainly large rivers, where only water quality data are collected and from which use assessments are made. The state has just embarked on an intensive watershed monitoring program in 1998, in which the first 5-year watershed cycle will concentrate primarily on a broad picture of water quality in the state. (VanArsdall, 2000)

In this watershed cycle, the state will sample approximately 350-400 random sites over the 5-year watershed cycle, concentrating on 1 to 3 major river basins each year. The watershed will be sampled for macroinvertebrates and habitat. These samples will allow the state to extrapolate aquatic life use to most miles of wadable streams from a 1:100,000 scale hydrologic network. (VanArsdall, 2000)

Kentucky does no random survey water quality sampling because of inadequate resources. For targeted water quality sampling, the fixed statewide network consists of 71 sites located at the downstream reaches of 8-digit cataloging units, mid-unit in the 8-digit watersheds, influent to major reservoirs, and major tributaries. These are sampled bimonthly except when they fall into the watershed cycle, and then they are sampled more frequently for that one year. In the rotating watershed water quality network, the state will sample about 30 sites each year that fill in the hydrologic gaps in the fixed network by picking up most of the 5th order waterhsheds. Some are also sited for other purposes such as predominant land use, TMDLs, least impacted, etc... Sampling frequency at these sites depends on the objective of the particular site. (VanArsdall, 2000)

Because of help from other federal and state agencies, Kentucky has much more biological sampling resources at their disposal, and these resources are used for targeted

biological sampling. They are able to sample most 4th order streams for at least one assemblage and habitat. This informs the state which basins have problems that need to be addressed by later sampling and mitigation activities. Over the 5-year watershed cycle, this targeted biological sampling will total over 1000 sites. (VanArsdall, 2000)

Alabama: This state follows the EPA recommended assessment criteria (percentages for chemical data). If there exists a large data set it is considered "monitored" data for assessment. "For example, 5 month (June-October), once-a-month sampling is considered *monitored,* but if the field personnel sample any less than this it would be considered *evaluated* data." Alabama is also developing specific site criteria for biological, physical/chemical, and habitat data, as well as criteria for determination of miles/acres affected. However, as of yet, Alabama does not have a state methodology for judging biology index/metrices results. (Reif, 2000)

Conclusions

This review indicates that many types of analyses are being used to provide information about water quality. The first major conclusion is that although there are some who criticize significance testing (Chapter II), this type of analysis is alive and well in the field of water quality. It is interesting to note that although hypothesis testing seems to be popular, as evidenced by its inclusion in guidance documents and water quality studies, the actual hypothesis tested is never reported, despite recommendations to the contrary in many of the guidance documents (Gilbert, 1987; Ward et al., 1990; Helsel and Hirsch, 1992; Montgomery and Reckhow, 1984; EPA, 1992; EPA, 1997c).

With a few exceptions (Heiskary, Lindbloom and Wilson, 1994; Momen et al., 1997; EPA, 1992; EPA, 1997c), the power of significance testing is not considered. The weight of evidence in making a decision about trends or differences in populations relies solely on the acceptable Type I error (α) and obtained significance level (p-value).

The literature review does not support the conclusion that there exist "de facto" standards for data analysis. The review of refereed journals found a large variety of graphical, statistical, and estimation analysis techniques. The EPA provides many types of guidance for different regulatory programs, yet the analysis recommendations differ between programs, and efforts do not seem to be coordinated between programs. It *was* apparent that specific methods were preferred by the USGS for trend detection (Seasonal Kendall test) and Differences in Populations (Wilcoxon Rank-Sum/Kruskal-Wallis and ANOVA).

The major commonalties to all the data analyses performed was that with a few exceptions: (1) justification was rarely given for choosing a certain test beyond the data being parametric or nonparametric, (2) the hypothesis tested was rarely stated, (3) alternative analysis methods, if explored, were not reported, and (4) the power of the significance test was never calculated.

Given the extremely wide array of data analysis methods being employed in producing information about water quality conditions, there is little reason to expect 'comparable' information is being produced in support of water quality management decision making. This fact leads to many of the criticisms highlighted in the previous chapter.

Chapter IV. Evaluation of Information Comparability Through Application of Different Data Analysis Methods

The previous chapters were dedicated to compilation of information in order to determine how water quality data are being analyzed for information purposes. Recent criticisms of statistical significance testing have questioned the main process through which information is produced from water quality data, i.e. significance testing. Nevertheless, the literature review established that using hypothesis testing is accepted in texts, guidance documents, and water quality studies published in refereed journals.

The literature review also establishes that there are a wide variety of methods that are available for data analysis. Many times, those who are analyzing water quality data are not statisticians, and rely on these texts, guidance documents, and observations of previous studies to select the analysis methods.

The purpose of this chapter is to document the connections between selection of data analysis methods and the comparability of the information produced. Using a 'high quality' data set provided by the New Zealand National Institute of Water & Atmospheric Research (NIWA), several different analysis methods were performed in the areas of trend detection, differences in populations, and standards compliance. The results of the different methods within each area were compared in order to illustrate how information changes depending on the analysis methods used.

Three statistical packages were utilized in the data analysis procedures. WQStat Plus™ (Version 1.5, developed by Intelligent Decision Technologies) was chosen for its

inclusion of nonparametric procedures, easy flow-adjustment and water quality data analysis focus. Minitab™ (Release 12, developed by Minitab Inc.) was chosen because of its broad base of statistical procedures, both parametric and nonparametric. MS-Excel™ (part of the Microsoft Office package) was also used for its basic statistical functions and ease of data manipulation (the data used was originally received in MS-Excel™ format). Comparison of results of like tests between statistical packages should also help to demonstrate the variability of information.

There are a large number of statistical packages that may be more commonly used for data analysis (i.e. S-Plus, SAS), but were not available for this research. It was hypothesized that results from different packages would be identical, and so no effort was made to acquire these packages prior to data analysis. This hypothesis will be discussed later in this chapter.

Approach for Demonstrating Various Statistical Methods on New Zealand Data Set

The New Zealand River Network data set was chosen for analysis because of its high quality and accessibility. The data record is from a 77 river-site monitoring network distributed throughout New Zealand's North and South Islands (Smith et al., 1996). The monitoring network's design is well documented and the network has been operated consistently over its 10-year life with excellent quality control procedures in place. The data was readily made available, in an easy to use format (MS-Excel™ Spreadsheets) for purposes of this study. (Refer to Appendix D for the actual data used in this study)

The format of the New Zealand data allowed for easy transition to data analysis, a reason that this particular set was chosen. The New Zealand data was accompanied by

meta-data that described the monitoring sites, how the samples were collected and analyzed, and all other ancillary data which would be of use to a data analyst (i.e. dates and units of measurement). Censored data (e.g. nondetects) were not used in this data, as all concentrations were reported. A few sites had missing data for certain dates, which were represented with a period (.) in the appropriate worksheet cell.

The only manipulation required for importation of the data into WQStat Plus™ and Minitab™, was cutting and pasting of the data columns into the appropriate format for the respective software. The required formats were described in the software user manuals (Intelligent Decision Technologies: p 34-50, 1998; Minitab: p 2-1 – 2-11, 1997).

A preliminary analysis for trends was performed after the first five years of monitoring, and results were published in a paper, *Trends in New Zealand's National River Water Quality Network* (Smith et al., 1996). This allowed comparison and verification of results of trend analysis for this thesis with results from Smith et al. (1996).

Selection of Three Sites and Constituents for Data Analysis

Not all sites or constituents of the River Network were analyzed as part of this demonstration. Sites and constituents were chosen upon review of the trends paper (Smith et al., 1996), and with input from Graham McBride, Project Director, NIWA, Hamilton, New Zealand. Descriptions of the sites were provided in the appendices of the New Zealand data set (Bryers, 1999; see Appendix D). For purposes of this study, four data records, at four sites, were selected as follows:

A. Site HM4 for BOD5 – This site is on the Waikato River, and is located downstream of the catchment area. It has potential impacts from agriculture, paper and pulp industries, and has additional inputs from Hamilton, Ngaruawahia, Huntly, thermal power stations, swamps, pasture and coal mining. (Bryers, 1999) The New Zealand Trends paper (Smith et al., 1996) showed no trend for BOD5 after the first 5 years at this site.

B. Site RO2 for NH4 analysis – This site is on the Tarawera River, a major river in the area, downstream of major pulp and paper industries and exotic forest plantations. There is agricultural pasture in the valley. (Bryers, 1999)
The New Zealand Trends paper (Smith et al., 1996) showed an upward trend in NH4 at ($p<5\%$) level for the first 5 years at this site.

C. Site RO1 for NH4 analysis – This site will only be used in the differences in population analysis. RO1 is upstream of site RO2 (above) on the Tarawera River. Between the two sites are potential environmental impacts from a pulp mill (The Tasman Pulp and Paper mill), farming, a town, Kawerau, and a geothermal area. (Bryers, 1999) This site was used as an upstream site for differences in population's analysis only.

D. Site HM6 for NO3 data – This site is not downstream of any urban sources, but is a major tributary of the Waihou River. It contains or will contain discharges from several large gold mining operations as well as agricultural

impacts from some pasture usage. (Bryers, 1999) The New Zealand Trends paper (Smith et al., 1996) showed an upward trend of NO3 at the ($p<5\%$) level after the first 5 years.

Testing Data for Normality

In order to illustrate the importance of distribution assumption in hypothesis testing, it was necessary to test each data set for normality. This was accomplished using the Chi-Squared Goodness-of-Fit Procedure in WQStat Plus™ (Intelligent Decision Technologies: p 71-72, 1998). In this procedure the calculated chi-square test statistic is compared to a table of chi-squared distributions with alpha = 0.05 and K-3 degrees of freedom, where K is the number of subgroups, or number of observations divided by an appropriate number (12 in this case). The null hypothesis as stated in WQStat Plus™ (Intelligent Decision Technologies: p 72, 1998) is:

Ho: the data are normally distributed (1)

vs.

Ha: the data are not normally distributed (2)

If the calculated value exceeds the tabulated value, then the program fails to reject the null hypothesis.

Flow Adjustment Procedures

Flow adjustment of the raw data was performed only in WQStat Plus™, as this was the only package that had the ability to directly calculate the flow-adjusted concentrations. This procedure was used to help determine how flow can affect or

change the information produced from the monitoring data. Flow adjusted concentrations (FAC) were used in normality testing, trend detection and standards compliance testing. The most common application of flow adjustment is trend analysis. For water quality constituents that are closely related to flow, an apparent trend in quality could be caused by a change in flow. By flow adjusting before trend analysis, the user can remove flow effects and determine the magnitude and statistical significance of trends that are not explained by flow. (Intelligent Decision Technologies p: 72, 1998)

WQStat Plus™ uses a log-log relationship assumption for its flow adjustment. The logs of the raw data are plotted against the logs of the flow. Then linear regression (least squares) is performed to determine the slope and the intercept of the line:

Log concentration = b*(log flow) + a

Then, from each water quality observation (concentration), the corresponding prediction based on flow, b(log flow) + a, is subtracted. This produces a flow-adjusted series of water quality observations with a sample mean of zero. To complete the adjustment, the overall sample mean of the water quality constituent series is added back in to each observation so that the mean of the flow-adjusted series is equal to the original mean. (Intelligent Decision Technologies p: 72 - 73, 1998)

It is realized that this procedure can introduce bias and error into the flow-adjusted data if the raw data does not fit the log-log model. However, the purpose of this procedure for WQStat Plus™ and this research is to give an indication of how flow-adjustment can change results from trend analysis.

Statistical Methods Used to Determine Trends

Analysis of the New Zealand data set for trends includes data from all ten years. As a means of additional quality control on the information being produced, analysis of the first 5 years was compared to the same analysis performed by a study published after the first 5 years of New Zealand's monitoring effort, entitled *Trends in New Zealand's National River Water Quality Network* (Smith et al., 1996). The second 5-year data was also analyzed separately, as well as a comparison of both 5-year analyses to an analysis of the 10-year data. Analyses were performed on raw data and flow-adjusted concentrations (FAC). The following statistical methods to detect trends were performed:

A. Mann-Kendall Test/Sen Slope Estimator – WQStat Plus™

The Mann-Kendall test for temporal trend is a nonparametric test, which uses the relative magnitude of the data, rather than actual values. The null hypothesis as stated in WQStat Plus™ (Intelligent Decision Technologies: p 77, 1998) is:

Ho: No significant trend of a constituent exists over time (3)

versus the alternative hypothesis:

Ha: A significant upward or downward trend exists over time (4)

In WQStat Plus™, a normal approximation was used because the New Zealand data set contained more than 41 points (sample size was approximately 120 for each set). A test statistic, Z, is computed and compared to a critical value, $Z_{1-\alpha/2}$ (for this two-tailed test). WQStat Plus™ tests for

trend at the significance levels (corresponding to acceptable Type I error) of α = 0.2, 0.1, 0.05 and 0.01 respectively. (Intelligent Decision Technologies: p 77-80, 1998) In this procedure, Sen's nonparametric slope estimator is also calculated. This is a nonparametric procedure used to estimate the true slope. (Intelligent Decision Technologies: p 81-82, 1998)

B. Seasonal Kendall Test – WQStat Plus™

The Seasonal Kendall Test is an extension of the Mann-Kendall Test that removes seasonal cycles and tests for trend. WQStat Plus™ uses the hypotheses listed above (equations (3) and (4)), and tests at the 80%, 90% and 95% confidence levels, which correspond to α = 0.2, 0.1, and 0.05 respectively. This procedure also includes a slope estimator. (Intelligent Decision Technologies: p 82-85, 1998)

The Seasonal Kendall Test was also used to test for trends in flow data at the three sites chosen: HM4, RO2 and HM6 for both 10-year data and each 5-year data set. This was performed to help in interpretation of the flow-adjusted trend results.

Statistical Methods Used to Determine Differences in Populations

The difference in population analysis was performed between the first 5-year and second 5-year data sets for the sites HM4, HM6 and RO2, as well as a test between sites RO1 and RO2 for NH4. The following tests, listed below, were performed for comparability of results. For further demonstration of comparability of results, the two-

sample t-test was performed in both MS-Excel™ and Minitab™, and the Mann-Whitney test was performed in WQStat Plus™ and Minitab™.

A. Two Sample T-test – MS-Excel™ and Minitab™

This is a standard parametric statistical test; perhaps the most widely used method for comparing two independent groups of data (Helsel and Hirsch, 1992). The t-test assumes that both groups of data are normally distributed around their means, and that they have the same variance. The null hypothesis for the two-sample t-test is stated in Minitab™ (Minitab™ Help, 1997) as:

Ho: $\mu_x = \mu_y$ the means for groups x and y are identical (5)

vs.

Ha: $\mu_x \neq \mu_y$ the means for groups x and y are not equal (6)

A two-tailed test was used in the New Zealand data analysis to avoid any assumptions of which group's mean might be higher. Helsel and Hirsch (1992) list five problems with the standard t-test that make it less applicable for general use than a nonparametric test. These are: 1) lack of power when applied to non normal data, 2) dependence on an additive model, 3) lack of applicability for censored data, 4) assumption that the mean is a good measure of central tendency for skewed data, and 5) difficulty in detecting non-normality and inequality of variance for the small sample sizes common to water resources data.

To help in interpreting data analysis results, the following analysis procedures were followed: First, each data set was tested for normality (see discussion above). Then, a standard two-sample F-test for equality of variances was applied using MS-Excel™. This test uses the F statistic and distribution to test the following null hypothesis:

Ho: $\sigma^2_x = \sigma^2_y$ The variances of two populations are equal (7)

vs.

Ha: $\sigma^2_x \neq \sigma^2_y$ The variances of the two populations are not equal(8)

The variances of site RO1 and RO2 for NH4 data rejected the null hypothesis, thus proven to be not equal, so the t-test for unequal variances was performed. All other data were analyzed with the two-sample t-test for equal variances. The only difference between the two t-tests is in modification of the degrees of freedom and t-statistic using Satterwaite's approximation. (Helsel and Hirsch: p126, 1992)

B. Mann-Whitney test – WQStat Plus™ and Minitab™

This is a nonparametric test for difference in populations. The null hypothesis tested in WQStat Plus™ (Intelligent Decision Technologies: p 95, 1998) is stated as:

Ho: The populations from which the two data sets have been drawn have the same mean. (9)

vs.

Ha: The populations have different means (10)

WQStat Plus™ uses a normal approximation for sample sizes > 10 for the test statistic calculation. It also tests and reports results for the 80%, 90%, 95% and 98% confidence levels (α of 0.2, 0.1, 0.05, and 0.02 respectively). (Intelligent Decision Technologies: p 95 - 98, 1998)

Minitab™ calculates the test statistic and the attained significance level (p-value), but it is not known if a normal approximation is used for large sample sizes. The main deviation from the WQStat Plus™ procedure is that the null hypothesis is stated as (Minitab™ Help, 1997):

Ho: the medians of two populations are equal (11)

vs.

Ha: the medians of the two populations are not equal (12)

C. Interval Tests – MS-Excel™

This is a parametric t-test procedure developed to test for differences in populations. Interval tests are largely used in the pharmaceutical industry involved in drug-testing analyses (Chow and Liu, 1992). The hypothesis for an interval test can take two forms, one testing for equivalence between groups, and one testing for inequivalence. Both the equivalence and inequivalence tests are used to determine whether the difference between means does not exceed an established interval. In the equivalence test, the null hypothesis (meaning the *tested* hypothesis, not implying that the difference is zero) is specifically:

WQStat Plus™ uses a normal approximation for sample sizes > 10 for the test statistic calculation. It also tests and reports results for the 80%, 90%, 95% and 98% confidence levels (α of 0.2, 0.1, 0.05, and 0.02 respectively). (Intelligent Decision Technologies: p 95 - 98, 1998)

Minitab™ calculates the test statistic and the attained significance level (p-value), but it is not known if a normal approximation is used for large sample sizes. The main deviation from the WQStat Plus™ procedure is that the null hypothesis is stated as (Minitab™ Help, 1997):

Ho: the medians of two populations are equal (11)

vs.

Ha: the medians of the two populations are not equal (12)

C. Interval Tests – MS-Excel™

This is a parametric t-test procedure developed to test for differences in populations. Interval tests are largely used in the pharmaceutical industry involved in drug-testing analyses (Chow and Liu, 1992). The hypothesis for an interval test can take two forms, one testing for equivalence between groups, and one testing for inequivalence. Both the equivalence and inequivalence tests are used to determine whether the difference between means does not exceed an established interval. In the equivalence test, the null hypothesis (meaning the *tested* hypothesis, not implying that the difference is zero) is specifically:

Ho: lower bound of equivalence interval <= μ_x - μ_y <= upper bound of equivalence interval (the difference in means lies within an accepted prior established interval) (McBride, 1999) (13)

The null hypothesis for an inequivalence test is:

Ho: μ_x - μ_y < lower bound of equivalence interval

Or μ_x - μ_y > upper bound of equivalence interval (the difference in means lies outside of an accepted prior established interval) (McBride, 1999) (14)

The difference between these two tests is that in the equivalence test, the assumption is that the populations are statistically and ecologically equivalent, whereas in the inequivalence test, the assumption is that the populations are not equivalent (a hypothesis which takes more precaution). Both tests recognize that the means will be different, but not necessarily equivalent. (McBride, 1999)

The interval chosen for these tests in this analysis was one of +/- 20% of the mean of the upstream or background data. While this was arbitrarily chosen, the estimates provided in McBride (1998) served as a guide for the magnitude. The purpose is to illustrate how different data analysis methods affect information. Establishing an equivalence interval requires knowledge of the behavior and affect of each constituent in the environment, something which is beyond the scope of this thesis.

A highly detailed explanation of the development of this type of testing used for environmental data can be found in McBride (1999a). The algorithm

through which the tests were performed in MS-Excel™ can be found in Appendix E (McBride, 1999b).

Statistical Methods Used to Determine Compliance (Standards Violations)

For these tests the New Zealand standard for BOD5 was compared to the data for BOD5 from site HM4. Although the country has few national numerical standards, 2 or 3 ppm is often the accepted limit set by waste load allocations (McBride, 1999c). The data set for site HM4 never exceeded 3 ppm, so for the purposes of this illustration, the excursion limit was set at 2 ppm. The following methods will be used:

A. Proportion Estimate – WQStat Plus™

This estimating procedure computes the proportion of observations in the record that exceed a stated excursion limit and computes a confidence limit for this proportion. In WQStat Plus™, the distribution model is the binomial distribution (success/failure distribution), and the significance levels reported are 95% and 99%. The proportions, upper and lower confidence limits are given for each season and the overall data set. (Intelligent Decision Technologies: p 98-100, 1998)

B. Tolerance Limits – WQStat Plus™

Tolerance limits define an interval that contains a specified fraction (coverage) of the population with specified probability (confidence level).

They are often used to compare concentrations from compliance stations to the upper limit of the tolerance interval. Calculations for this procedure are provided in the WQStat Plus™ user manual (Intelligent Decision Technologies: p 100-103, 1998)

For the tolerance limit procedure, an interval was established with 95% coverage from the 1st five years of data (background), and then the upper limit of the interval was compared to the 2nd 5 years (compliance) data. If compliance concentrations fall above the upper limit of the tolerance interval, this provides statistical evidence of a difference (Intelligent Decision Technologies, 1998). If more than 1-α fall outside the limits (5%) the evidence of a difference is statistically significant. However any excursion of the limit might indicate further need for investigation. (Intelligent Decision Technologies, 1998) Both parametric and nonparametric estimating procedures were performed for comparison.

C. Tolerance Interval – WQStat Plus™

Like the Tolerance Limit procedure, the Tolerance Interval estimation procedure is defined by tolerance limits for a specified coverage and confidence level. However, in the tolerance interval procedure, an interval was established from all of the data (instead of compliance data), which contained 95% coverage at the 95% confidence level. This interval was compared to the excursion limit of 2 ppm (instead of limit determined by background data). This estimating procedure was performed both

parametrically and nonparametrically. For a complete explanation of the calculations, see the WQStat Plus™ user manual (Intelligent Decision Technologies: p 108-110, 1998).

D. Confidence Interval – WQStat Plus™

This estimation interval is constructed with a mean concentration (parametric procedure) or a median concentration (nonparametric procedure) with a designated level of confidence. If the entire confidence interval exceeds the compliance limit, this is statistically significant evidence that the mean concentration exceeds the compliance limit. (Intelligent Decision Technologies: p 105-108, 1998) Both the parametric and nonparametric procedures were used.

E. Prediction Limits – WQStat Plus™

The prediction limit method used the 1^{st} 5-year data as background to establish an interval, and the 2^{nd} 5-year data were compared to the interval to determine excursions. The interval includes k future observations from the same population with a specified confidence (95%). If any observation exceeds the bounds of the prediction interval, this is statistically significant evidence that the observation is not representative of the background group. (Intelligent Decision Technologies, 1998) If there is more than one source of variation, the parametric Prediction Limit should is inappropriate. The complete procedure can be found in the WQStat Plus user manual™

(Intelligent Decision Technologies: p 103-105, 1998). Both parametric and nonparametric methods were used for comparison.

Results of Data Analysis

The following section examines the results of applying the methods discussed above. Particular attention is paid to comparing the differences in results (i.e. information) that are consequences of changing the analysis method. It is the lack of comparable information resulting from arbitrary selection of data analysis methods that is the focus of the results presentation.

Testing for Normality

All data sets were tested for normality in order to interpret the resulting information from parametric and nonparametric significance tests. This was accomplished through the Chi-Square Goodness of fit test in WQStat Plus™, in which the null hypothesis is that the data are normally distributed (stated in equation (1)).

Table IV.1: Normality Testing Results

Site Constituent	Hypothesis Test Result	Conclusion
RO1_NH4 (raw)	Reject the null hypothesis	Not normal
RO2_NH4 (raw)	Fail to reject the null	Cannot prove normal
RO2_NH4 (FAC)	Fail to reject the null	Cannot prove normal
HM4_BOD5 (raw)	Reject the null hypothesis	Not normal
HM4_BOD5 (FAC)	Fail to reject the null	Cannot prove normal
HM6_NO3 (raw)	Fail to reject the null	Cannot prove normal
HM6_NO3 (FAC)	Fail to reject the null	Cannot prove normal

Comments

Raw vs. flow-adjusted concentrations (FAC) affected the outcome of this test. Most data sets tested failed to reject the null hypothesis that they were normally distributed. However, as discussed in the Background Chapter, failure to reject a null hypothesis does not prove that it is true. This is why there is a question as to whether these data are normally distributed or not. This test can only give confidence (95%) that a data set is *not* normally distributed. (See Appendix F for WQStat Plus™ results)

Results for Trend Detection

This analysis compared the Mann-Kendall/Sen's Slope Estimator (MK) for trend with the Seasonal Kendall (SKT) test on 10-year data, raw and flow-adjusted (FAC), as well as the 1st and 2nd 5-year data. All calculations were performed using WQStat Plus™.

Table IV.2: Trend Detection Results for Site HM6, Constituent NO3

Data	Test	Results	Slope Estimate
10 yr – flow	*SKT*	*⇓ - 80% Confidence Level*	*-0.1073 units/year*
10 yr – raw	MK	Fail to reject null of no trend	2.955 units/year
10 yr – raw	SKT	Fail to reject null of no trend	1.929 units/year
10 yr - FAC	MK	⇑ - 95% Confidence Level	11.125 units/year
10 yr - FAC	SKT	⇑ - 90% Confidence Level	8.953 units/year
1st 5 yr – flow	*SKT*	*⇓ - 95% Confidence Level*	*-0.8778 units/year*
1st 5 yr – raw	MK	Fail to reject null of no trend	-9.359 units/year
1st 5 yr – raw	SKT	⇓ - 80% Confidence Level	-28.25 units/year
1st 5 yr - FAC	MK	⇑ - 95% Confidence Level	36.81 units/year
1st 5 yr - FAC	SKT	⇑ - 90% Confidence level	28.42 units/year
2nd 5 yr – flow	*SKT*	*Fail to reject null of no trend*	*-0.1298 units/year*
2nd 5 yr - raw	MK	Fail to reject null of no trend	7.953 units/year
2nd 5 yr - raw	SKT	Fail to reject null of no trend	20.13 units/year
2nd 5 yr - FAC	MK	⇑ - 90% Confidence Level	27.24 units/year
2nd 5 yr - FAC	SKT	⇑ - 80% Confidence Level	23.28 units/year

Results

Both tests showed no significant trend at any alpha (α) or confidence for the 10-year raw data, but detected an upward trend in the 10-year flow-adjusted concentrations. Mann-Kendall detected at the 0.05 α, and Seasonal-Kendall at the 90% confidence (α = 0.1).

The Mann-Kendall test resulted in no trend for the 1st 5-year raw data, the Seasonal Kendall test detected a downward trend at the 80% confidence level (α = 0.2). When flow-adjusted concentrations were used, both tests showed an upward trend, Mann-Kendall at α = 0.05, Seasonal Kendall at α = 0.1.

No significant trend was found for the 2nd 5-year raw data. Mann-Kendall showed an upward trend in flow-adjusted concentrations at an α = 0.1, and Seasonal Kendall showed an upward trend at α = 0.2.

The trend results on flow (see Appendix K for results) showed a downward trend in flow for the 10-yr data at the 80% confidence level (α = 0.2). The first 5-yr data showed a downward trend at 95% confidence, but the second 5-yr data failed to reject the null of no significant trend.

Comments

Findings are similar for both tests, but not exact. It is often standard practice to choose an acceptable Type I error of 0.05 (95% Confidence Level). If that were the case in this analysis, only the Mann-Kendall test would have detected any trends in the 10-year flow-adjusted concentrations and the 1st 5-year flow-adjusted concentrations. WQStat™ gives results for various alphas (confidence levels) up to 0.2 (80% confidence)

and so allows the user to see the alpha giving a significant result. These results illustrate that findings can change by choosing a confidence level (α) after results are obtained.

Flow-adjusted concentrations changed the outcome of the trend test upon examination of the trendline in the time series plot and in the 1st 5-year significance test, as the direction changed from downward to upward trend. The slope estimators seem to have similar (i.e. comparable) results. (See Appendix G for Trend Analysis results) It is interesting to note that where a downward trend in flow existed, so did an upward trend in constituent concentration in flow-adjusted concentrations, but not exclusively. This finding could aid in the interpretation of the temporal behavior of the constituent.

Table IV.3: Trend Detection Results for Site HM4, Constituent BOD5

Data	Test	Results	Slope Estimate
10 yr – flow	*SKT*	*Fail to reject null of no trend*	*-4.692 units/year*
10 yr – raw	MK	⇓ - 99% Confidence Level	-0.033 units/year
10 yr – raw	SKT	⇓ - 95% Confidence Level	-0.0332 units/year
10 yr - FAC	MK	⇓ - 99% Confidence Level	-0.034 units/year
10 yr - FAC	SKT	⇓ - 95% Confidence Level	-0.03591 units/year
1st 5 yr – flow	*SKT*	*⇓ - 95% Confidence Level*	*-45.21 units/year*
1st 5 yr – raw	MK	Fail to reject null of no trend	-0.016 units/year
1st 5 yr – raw	SKT	Fail to reject null of no trend	0 units/year
1st 5 yr - FAC	MK	Fail to reject null of no trend	-0.039 units/year
1st 5 yr - FAC	SKT	Fail to reject null of no trend	-0.03132 units/year
2nd 5 yr – flow	*SKT*	*Fail to reject null of no trend*	*1.211 units/year*
2nd 5 yr - raw	MK	Fail to reject null of no trend	-0.028 units/year
2nd 5 yr - raw	SKT	Fail to reject null of no trend	-0.04568 units/year
2nd 5 yr - FAC	MK	Fail to reject null of no trend	-0.025 units/year
2nd 5 yr - FAC	SKT	Fail to reject null of no trend	-0.04418 units/year

Results

Both tests give a significant downward trend in 10-year raw and flow-adjusted concentration data at all alphas. First 5-year raw and flow-adjusted concentration data

show no trend for both tests at all alpha levels. Second 5-year raw and flow-adjusted concentration data show no significant trend for both tests at all alpha levels. There was a large downward trend in flow in the first 5-year data, but no significant trend in the 10-year or second 5-year data (See Appendix K for results).

Comments

These findings illustrate how significance tests are more likely to detect a trend as sample size increases, a phenomenon common to all the tests performed in this chapter. No trend was detected in either 5 years of data, but was detected in the 10-year data. These results were determined by comparing a calculated test statistic to a tabled value, and not by a calculated p-value (observed significance level). Therefore, the results from the five-year tests are comparable to the ten-year tests, although the sample sizes are different (see discussion in Chapter V). The slope estimates are highly comparable at this site. (See Appendix G for complete results) Determination of flow trend did not reveal anything about flow-adjusted constituent behavior.

Table IV.4: Trend Detection Results for Site RO2, Constituent NH4

Data	Test	Results	Slope Estimate
10 yr – flow	*SKT*	*Fail to reject null of no trend*	*0.01263 units/year*
10 yr – raw	MK	⇑ - 90% Confidence Level	1.142 units/year
10 yr – raw	SKT	⇑ - 95% Confidence Level	1.283 units/year
10 yr - FAC	MK	⇑ - 95% Confidence Level	1.391 units/year
10 yr - FAC	SKT	⇑ - 95% Confidence Level	1.344 units/year
1st 5 yr – flow	*SKT*	*⇓ - 95% Confidence Level*	*-2.105 units/year*
1st 5 yr – raw	MK	⇑ - 99% Confidence Level	7.063 units/year
1st 5 yr – raw	SKT	⇑ - 95% Confidence Level	6.53 units/year
1st 5 yr - FAC	MK	⇑ - 99% Confidence Level	6.044 units/year
1st 5 yr - FAC	SKT	⇑ - 95% Confidence Level	5.305 units/year
2nd 5 yr – flow	*SKT*	*⇑ - 95% Confidence Level*	*1.47 units/year*
2nd 5 yr - raw	MK	⇓ - 99% Confidence Level	-4.991 units/year
2nd 5 yr - raw	SKT	⇓ - 95% Confidence Level	-4.26 units/year
2nd 5 yr - FAC	MK	⇓ - 95% Confidence Level	-4.219 units/year
2nd 5 yr - FAC	SKT	⇓ - 90% Confidence Level	-3.066 units/year

Results

10-year raw data shows an upward trend at $\alpha = 0.1$ (90% confidence level) for the Mann-Kendall test, and $\alpha = 0.05$ (95% confidence level) for the Seasonal Kendall test. Flow-adjusted concentrations show an upward trend at the 95% confidence level for both tests. The first 5-year raw and flow-adjusted concentration data show an upward trend for both tests at all alpha levels. The second 5-year raw data show a downward trend for both tests at all alpha levels. Flow-adjusted concentrations give a downward trend at $\alpha = 0.05$ (95 % confidence level) for the Mann-Kendall test, and $\alpha = 0.1$ (90% confidence level) for the Seasonal Kendall test.

Flow trend results failed to reject the null of no trend for the 10-year data, but showed a downward trend in the first 5-year data (95% confidence) and an upward trend in flow for the second 5-year data (95% confidence) (See Appendix K).

Comments

These findings illustrate how an upward trend in the first half of the constituent data record and a downward trend in the second half of the record might reconcile itself. The upward trend was stronger than the downward trend, and so was detected in the overall 10-year data. Again, the level of detection of trend was different for both tests. However, this time the Seasonal Kendall test was more sensitive in the 10-year HM4 data set, as opposed to results for site HM6, in which the Mann-Kendall test seemed more sensitive. The Mann-Kendall test detected a trend at a smaller alpha level in the 2nd 5-year flow-adjusted concentrations. The slope estimates are very comparable at this site. (See Appendix G for complete results) Again, a downward trend in flow correlated to an upward constituent trend, and vice versa, in both the raw and flow-adjusted constituent concentrations.

Results for Differences in Populations Analysis

This series of analyses compared the first 5-year data to the second 5-year data for BOD5 (site HM4), NO3 (site HM6) and NH4 (site RO2). This is often referred to as step trend detection, but in actuality is a test for population differences before and after a specific point in time. To illustrate an analysis for spatial differences, a comparison was made between upstream and downstream NH4 values for sites RO1 (u) and RO2 (d).

These analyses utilized the nonparametric Mann-Whitney test in Minitab™ and WQStat Plus™, and the two-sample t-test in MS-Excel™ and Minitab™. Differences were also sought through Interval tests developed in MS-Excel™ by Graham McBride

(See Methods section above, and Appendix E). The t-test for equal variances was used in all cases except for site RO1 vs. RO2. As described in the Methods section, results of F-tests for equal variances in MS-Excel™ resulted in the finding of equal variances between the first and second 5-year data from each site, as expected. However, the F-test resulted in the finding of unequal variances between RO1 and RO2 data for NH4, therefore requiring the use of the t-test for non-equal variances. (For F-test results see Appendix H).

Table IV.5: Differences in Population Results for Site HM4, Constituent BOD5

Test	Results
MS-Excel™ t-test (1^{st} 5-yrs vs. 2^{nd} 5-yrs)	Significant Difference (p = 0.019)
Minitab™ t-test (1^{st} 5-yrs vs. 2^{nd} 5-yrs)	Significant Difference (p = 0.019)
Equivalence Interval test	Equivalent (2^{nd} 5-yrs within interval of +/- 20% of 1^{st} 5-yrs mean – 95% confidence)
Inequivalence Interval test	Inequivalent (2^{nd} 5-yrs *not* within interval of +/- 20% of 1^{st} 5-yrs mean – 95% confidence)
Minitab™ Mann-Whitney	Significant difference (p=0.0148)
WQStat™ Plus Mann-Whitney	Shows no significant difference (see below)

Results

Two-sample t-tests (two-tailed assuming equal variances) in MS-Excel™, and Minitab™ gave identical results of a significant difference in BOD5 between the first and second 5-year data (p = 0.019).

The interval test for equivalence failed to reject the null of equivalent mean concentrations (see equation (13)) at an $\alpha = 0.05$ and equivalence interval of +/- 20% change in the first 5-year mean. However, when the null hypothesis for the equivalence t-test is changed to inequivalence, the result is the failure to reject the null of inequivalent

concentrations (equation (14)) in population BOD5 at alpha = 0.05 and equivalence interval of +/- 20% change in the first 5-year mean.

Using the nonparametric Mann-Whitney test, Minitab™ gave a significant difference in BOD5 at p = 0.0148. WQStat Plus™ showed no rejection of the null of equal means at all confidence levels, though the test statistic calculated should have rejected the null hypothesis and found a significant difference.

Comments

These findings vary depending on alpha level, test and hypothesis. This illustrates how important assumptions of distribution and hypothesis are when testing, as well as selection of an acceptable Type I error (α). Again it illustrates that choosing the confidence level needed (α) after results are obtained can change the information obtained.

Minitab™ and WQStat Plus™ gave comparable results for the Mann-Whitney test, however a mistake in the WQStat Plus™ software misinterpreted the final results. In general the results from different statistical packages are comparable, though results are presented differently in each one.

At the beginning of this section it was found that the raw data for site HM4_BOD5 are not normally distributed. This could mean that a parametric t-test is not appropriate, as a nonparametric procedure could be more powerful. Therefore, the best information from this analysis comes from the Mann-Whitney test. (See Appendix I for Differences in Populations results)

Table IV.6: Differences in Population Results for Site HM6, Constituent NO3

Test	Result
MS-Excel™ t-test (1st 5-yrs vs. 2nd 5-yrs)	Fail to reject the null of equal means
Minitab™ t-test (1st 5-yrs vs. 2nd 5-yrs)	Fail to reject the null of equal means
Equivalence Interval test	Fail to reject the null of equivalence
Inequivalence Interval test	Rejected the null of inequivalence (2nd 5-yrs *within* interval of +/- 20% of 1st 5-yrs mean – 95% confidence)
Minitab™ Mann-Whitney	Fail to reject the null of equal medians
WQStat™ Plus Mann-Whitney	Fail to reject the null of equal means

Results

Two-sample t-tests (two-tailed assuming equal variances) in MS-Excel™ and Minitab™ failed to reject the null of equal means (equation (5)) between the first and second 5-year data ($p = 0.51$).

The interval test with either null hypothesis of equivalence (equation (13)) or inequivalence (equation (14)) resulted in equivalent populations at 95% confidence (alpha = 0.05) and an equivalence interval of +/- 20% of the 1st 5-year mean.

In computing the Mann-Whitney test, both Minitab™ and WQStat Plus™ failed to reject the null of equal medians (equation (11)) or means (equation (9)) between groups at $\alpha = 0.1$.

Comments

All of these tests failed to reject the null hypotheses of equal central tendency between the first and second 5-year data. This data also failed to reject the null of normal distribution, so the t-tests are more powerful tests of the difference in the two populations. However, failure to reject the null of equal means in the standard t-test does

not prove that they are equal. The best information in this analysis comes from the equivalence test with the null hypothesis that the two populations are *inequivalent* (equation (14)). Rejection of this null proves with 95% confidence that the mean of the second 5-year data lays within an interval of +/- 20% of the first 5-year data mean, making them equivalent. Of course, this is supposing that the +/- 20% change is an ecologically acceptable change in NO3. (See Appendix I for complete results)

Table IV.7: Differences in Population Results for Site RO2, Constituent NH4

Test	Result
MS-Excel™ t-test (1^{st} 5-yrs vs. 2^{nd} 5-yrs)	Fail to reject the null of equal means
Minitab™ t-test (1^{st} 5-yrs vs. 2^{nd} 5-yrs)	Fail to reject the null of equal means
Equivalence Interval test	Fail to reject the null of equivalence
Inequivalence Interval test	Rejected the null of inequivalence (2^{nd} 5-yrs *within* interval of +/- 20% of 1^{st} 5-yrs mean – 95% confidence)
Minitab™ Mann-Whitney	Fail to reject the null of equal medians (p=0.259)
WQStat™ Plus Mann-Whitney	Fail to reject the null of equal means

Results

Two-sample t-tests (two-tailed assuming equal variances) in MS-Excel™ and Minitab™ failed to reject the null of equal means (equation (5)) between the first and second 5-year NH4 data ($p = 0.18$).

The Interval test with a null hypothesis of equivalent means (equation (13)) failed to reject the null, whereas the Interval test with a null hypothesis of inequivalence (equation (14)) rejected the null of inequivalent means at 95% confidence ($\alpha = 0.05$) and an equivalence interval of +/- 20% of the first 5-year mean.

Calculating the Mann-Whitney test statistic in Minitab and WQStat Plus failed to reject the null hypotheses of equal medians (equation (11)) or means (equation (9)) at the 90% confidence level ($\alpha = 0.1$).

Comments

This NH4 data failed to reject the null of normal distribution, so the t-test is an appropriate and powerful test. However, as in the analysis at the previous site (HM6), failure to reject the null of equal means does not prove that the means are in fact exactly equal. Again the best information comes from the equivalence test with the null hypothesis that the two populations are *inequivalent* (equation (14)). Rejection of this null proves with 95% confidence that the mean of the second 5-year NH4 data lies within an interval of +/- 20% of the first 5-year NH4 data mean, making them equivalent. (See Appendix I for complete results)

Table IV.8: Analysis of Differences Between NH4 at RO1 and RO2

Test	Result
MS-Excel™ t-test (1^{st} 5-yrs vs. 2^{nd} 5-yrs)	Significant Difference (p=0.000)
Minitab™ t-test (1^{st} 5-yrs vs. 2^{nd} 5-yrs)	Significant Difference (p=0.000)
Equivalence Interval test	Rejected the null of equivalence (RO2 *not* within interval of +/- 20% of RO1 – 95% confidence)
Inequivalence Interval test	Fail to reject the null of inequivalence
Minitab™ Mann-Whitney	Significant Difference (p=0.000)
WQStat™ Plus Mann-Whitney	Significant Difference-99% confidence

Results

Two-sample t-tests (two-tailed assuming *unequal* variances, as discussed above) in MS-Excel™ and Minitab™ result in a significant difference between the means of NH4 at sites RO1 and RO2 ($p = 0.000$).

The Interval tests with both hypotheses of equivalence and inequivalence support significant differences in concentration of NH4 at 95% confidence (alpha = 0.05) and an equivalence interval of +/- 20% of the upstream (RO1) mean concentration.

The Mann-Whitney test in both Minitab™ and WQStat Plus™ result in significant differences between the medians/means of NH4 at sites RO1 and RO2.

Comments

This analysis shows that when the concentration differences are large between populations, distribution assumptions, hypotheses and alphas do not have a great affect on the results. Although NH4 at site RO2 failed to reject the null of normal distribution, the Mann-Whitney test is most appropriate because NH4 at site RO1 is not normally distributed. (See Appendix I for complete results)

Results for Standards Compliance

Standards compliance analysis alternatives were examined using the BOD5 data for site HM4. Common limits in New Zealand are 2 or 3 ppm. The analyses were performed using 2 ppm, since no data exceeded the 3 ppm limit. Excursion analysis was performed on both raw and flow-adjusted concentrations. The following analyses were

performed in WQStat Plus™: Proportion Estimates, Tolerance Limits, Tolerance Intervals, Prediction Limits, and Confidence Intervals about the mean.

Table IV.9: Standards Compliance Results for Site HM4, Constituent BOD5

Test	Compliance Results
Proportion Estimate – raw	3.3% excursions (0,7%) CI
Proportion Estimate – FAC	3.3% excursions (0,7%) CI
Parametric Tolerance Limit – raw	Exceeded limit
Parametric Tolerance Limit – FAC	Exceeded limit
Nonparametric Tolerance Limit – raw	Compliant
Nonparametric Tolerance Limit – FAC	Exceeded limit
Parametric Tolerance Interval – raw	Compliant
Parametric Tolerance Interval – FAC	Compliant
Nonparametric Tolerance Interval – raw	Exceeded limit
Nonparametric Tolerance Interval – FAC	Exceeded limit
Parametric Prediction Limit – raw	Exceeded limit
Parametric Prediction Limit – FAC	Compliant
Nonparametric Prediction Limit – raw	Compliant
Nonparametric Prediction Limit – FAC	Exceeded limit
Parametric Confidence Interval for the mean - raw	Compliant
Parametric Confidence Interval for the mean - FAC	Compliant
Nonparametric Confidence Interval for the median - raw	Compliant
Nonparametric Confidence Interval for the median - FAC	Compliant

Results

For Standards Compliance results from WQStat Plus™, see Appendix J. Both the raw and flow-adjusted concentrations data gave a 0.033 (3.3%) excursion proportion, with the 95% confidence interval ranging from 0 to 7% excursions.

The Tolerance Limit procedure was performed using the first 5-year BOD5 data establishing the Tolerance Interval. Then the second 5-year BOD5 data were compared to that interval for compliance. Both the raw and flow-adjusted concentration data exceeded the limit in the parametric Tolerance Limit procedure, but only the flow-

adjusted BOD5 concentrations exceeded the Tolerance Limit in the nonparametric procedure.

For the Tolerance Interval procedure, the compliance limit (2 ppm) is used to determine the excursion, not the background data (first 5-year data, as discussed in the Statistical Methods section). In this analysis, neither the raw nor flow-adjusted concentration data exceeded the parametric Tolerance Limit (95% coverage). However, both exceeded the nonparametric procedure limit.

In the parametric Prediction Limit procedure, the raw BOD5 data exceeded the Prediction Limit, whereas the flow-adjusted concentration data did not. For the nonparametric procedure, the opposite was true. The raw data did *not* exceed its Prediction Limit, whereas the flow-adjusted concentration data did.

In both the parametric and nonparametric Confidence Interval determinations, neither the raw nor flow-adjusted concentrations data means/medians exceeded the excursion limit of 2 ppm.

Comments

Each of these analyses gives different kinds of information about the data. The most straightforward is the proportion estimate, which tells exactly the proportion of excursion, along with a confidence interval so that the data can be representative of not only the sample, but also the population as a whole. These findings show that 3.3% of the data exceeded the excursion, and that up to 7% exceedance can be expected at the 95% confidence level.

The other procedure's outcomes (Tolerance Limit, Tolerance Interval, Prediction Limit and Confidence Interval) were highly influenced by the distribution assumption, and the concentration used (raw vs. flow-adjusted concentrations). The raw BOD5 data was shown to be not normal in the Testing for Normality section, so the nonparametric results are more appropriate in assessing compliance. The Tolerance Limit/Interval and Prediction Limit procedures are more appropriate for determining if a single sample exceeds a compliance limit or interval based on background data. Whereas the Confidence Interval is more appropriate for determining if the mean/median of a population exceeds a standard that is based on central tendency. The variety of results again illustrates the noncomparability of information produced from different analysis methods.

Chapter V. Discussion

The previous chapters have established that: (1) there are a large variety of methods employed in water quality data analysis to produce information; (2) significance testing is by far the most popular type of analysis used to interpret water quality monitoring data (used in 17 of 19 Trend Studies and 16 of 20 Differences in Population Studies from Chapter III), and; (3) many of these common methods, when applied to one set of data, do not produce comparable results.

When completing a water quality assessment, it is usually assumed that the analyst will make an independent decision based on his or her interpretation of the data and information needs, after the data are collected. This fact introduces considerable uncertainty into the analysis of water quality data and results in non-comparable information. This raises concerns about the actual management decision, stemming from the information on which it was based. If there is a lack of confidence in the methods used to produce information for management, then there will be a lack of confidence in the ultimate decision as well. The only way to instill confidence in the management decision is to remove the concerns over the process through which information for the decision was created.

'Standard' Data Analysis Methods?

This issue raises the question: Is it feasible to develop a set of 'standard' water quality data analysis methods for specific forms of management information (i.e. trends, differences, standards compliance) that can produce comparable information that is defensible? The simple answer is yes, as this question is not new to water quality management. "Perhaps the best way to ensure that data collected during different studies are comparable is to encourage all investigators to use standardized sampling and analysis protocols whenever possible " (Becker and Armstrong, 1988). Currently there are professionals in the field who have been charged with determining which sampling and laboratory analysis methods result in comparable information (see Methods and Data Comparability Board of the National Water Quality Monitoring Council; http://wi.water.usgs.gov/pmethods). This is an especially pertinent issue as the interest in data sharing continues to rise.

This suggestion is not made without reservation. A natural conflict stems from the need to obtain comparable information, and permitting site-specific conditions to be considered in how data are analyzed and interpreted. The answer to this issue is not readily apparent, nor are professionals studying the problem and its solutions. At present, the discussions of 'appropriate' use of statistics in water quality monitoring tend to be within various water-management related agencies. The literature review in Chapter III clearly illustrates that some agencies have produced guidance for data analysis over the years, yet without much coordination within or outside of the agency. The National Water Quality Monitoring Council is currently facing the issue described here, and

exploring the mechanisms that could help monitoring systems produce comparable information.

Several issues besides the methods selection itself will need to be addressed. Although some advise to the contrary (Ward et al., 1986), many analysts select the analysis methods after examining the data and its distribution. In fact, this is recommended by existing guidance (i.e. Montgomery and Reckhow, 1984; Chatfield, 1985). Chatfield (1985) recommends the following process: (1) Clarify the objectives of the investigation; (2) Collect the data in an appropriate way; (3) Investigate the structure and quality of the data; (4) Carry out an initial examination of the data; (5) Select and carry out an appropriate formal statistical analysis; (6) Compare the findings with previous results or acquire further data if necessary; and (7) Interpret and communicate the results." If 'standard' data analysis methods are developed, should they follow this same line of thinking?

There are good arguments for both sides of this issue. Choosing the analysis method before examining the data allows for impartial agreement and approval of the process by all interested parties without the bias of data results. However, choosing the method after analysis allows for selection of the most scientifically appropriate methods for the type of data gathered, without prior assumptions, but also allows for post-hoc selection of alpha, which, as illustrated in Chapter IV, can greatly influence the results. This issue in and of itself begs the assistance of professionals who are knowledgeable about water management to provide guidance for data analysis protocols.

Another topic that develops from the suggestion of standardizing data analysis methods deals with the extent that the analyst is allowed to produce information that

directly relates to the management decision-making. Most management decision-makers are not statisticians. Should results of analysis only be presented (such as a rejection of a null hypothesis and obtained p-value), or an interpretation in terms of meaning presented as well? Should management be allowed to decipher statistical results, without the bias of the analyst? Guidance is needed for these questions to be resolved. Only those involved in water management know the expertise of their colleagues in understanding these scientific issues. Comprehension will vary among managers, and so may the role of the analyst in interpreting information produced from the data analysis. The EPA (1998) dealt with this issue in the development of their *Guidelines for Ecological Risk Assessment*. The following process was recommended: "To ensure mutual understanding between risk assessor [i.e. analysts] and managers, a good risk characterization will express results clearly, articulate major assumptions and uncertainties, identify reasonable alternative interpretations, and separate scientific conclusions from policy judgments. Risk managers use risk assessment results, along with other factors (e.g. economic or other legal concerns), in making risk management decisions and as a basis for communicating risks to interested parties and the general public."

Finally, the question that directly pertains to the work presented in this thesis is: What would these 'standard' data analysis methods look like? With the exception of a few estimation and graphical procedures, the methods used in the previous chapter were all based on the statistical theory of significance testing, which Chapter II established is "under fire" in some parts of the scientific world. It is easy to see in the results of the New Zealand data analysis (Chapter IV) that information changes depending on the

method selection, but why? The answer lies in several flaws of applying significance testing to environmental (observational) data.

One flaw, which is rarely understood, is that results based on p-values from tests with different sample sizes are not comparable. A calculated p-value is affected not only by the data collected, but also by the data which might have been observed if the trial had gone differently than it in fact did (DuPont, 1983). Therefore, premature termination of an experiment (or monitoring effort) affects the outcome of the final calculated p-value. Unfortunately, there is often no way of knowing whether a test was performed at the end of an experiment, or in the middle, and so reported p-values might not be comparable, even for similar sample sizes.

The greatest of these flaws, which has been mentioned previously, is that the resource managers and analysts of water quality monitoring data are often not statisticians, and so are repeatedly guilty of choosing analysis methods without a thorough understanding of the underlying assumptions, meaning of test parameters, or interpretation of results. Johnson (1999) states, "While many of the arguments against significance tests stem from their misuse, rather than intrinsic value, I believe that one of their intrinsic problems is that they encourage misuse".

Why Use Significance Testing?

Nester (1996) suggests several reasons why hypothesis tests are so widely used: (1) they appear to be objective and exact; (2) they are readily available and easily invoked in many commercial statistics packages; (3) everyone else seems to use them; (4) students, statisticians and scientists are taught to use them; and (5) some journals and

editors and thesis supervisors demand them. The research in the previous chapters validates these claims. Yet the best explanation of why hypothesis testing is so popular rests on the foundation of the scientific method. Under that method, a theory is postulated, which generates predictions, or hypotheses. A scientific experiment is conducted to 'test' the hypothesis. The results of the experiment either refute the hypothesis, dictating that the theory is incorrect, or do not refute the hypothesis, letting the theory stand. In contrast, statistical hypotheses employed by environmental data analysts are known a priori to be false (Johnson, 1999).

So why test statistical hypotheses at all? McBride (2000) states that comparison of p-values for tests with similar numbers of samples does provide an elegant way of ranking the importance of differences measured, if sample sizes are identical. He also acknowledges that in constructing models, p-values are most useful in determining important explanatory variables in statistical models. However, this is more a function of exploratory data analysis, and not data analysis that better connects water quality information to management decision-making.

One answer would be that a statistical test could be only one factor in evidence of interpretation of the data. In this way, a single rejection of a point null hypothesis, or a p-value, would not be the only information leading to a management decision. Other pieces of information would need to be gathered to either support or refute the findings of the statistical test. EPA (1998) has produced guidance for ecological risk assessment that follows this type of process.

"Ecological risk assessment evaluates the likelihood that adverse ecological effect may occur or are occurring as a result of exposure to one or more stressors. It is a

flexible process for organizing and analyzing data, information, assumptions and uncertainties. Ecological risk assessment provides a critical element for environmental decision making by giving risk managers an approach for considering available scientific information along with the other factors they need to consider (e.g. social, political, legal or economic), in selecting a course of action." (EPA, 1998)

There exist alternatives to statistical testing which can provide scientifically defensible information to management about the quality of the water being monitored. It is not within the scope of this thesis to provide great detail about analysis alternatives, but the following section will outline some of the other pieces of information that could accompany or even replace statistical tests in order to make the information more comparable and meaningful to management.

Data Analysis Tools to Make Information More Comparable

There are many procedures that can be applied along with statistical tests in order to give more meaning to the results beyond the p-value. It might be assumed that these procedures are already mandatory for statistical analysis of water quality data, yet the literature review in Chapter III suggests that they are not. The first of these is to test the data for normality, and if the data are not normal, only use nonparametric analysis procedures, which are more powerful than parametric procedures for non-normal data, being less affected by nondetects or extreme values. The second is to use flow-adjusted concentrations, especially for trend detection and standards compliance. The third is to consider the power of the test. This gives a good indication of the likelihood of actually detecting an effect of the size practical to the analyst.

Power Analysis

Power analysis is becoming more prevalent due to the availability of statistical software packages and Internet "power calculators". However, the increase in availability does not directly translate into an increase in calculating the true power. Cursory exploration of three Internet calculators (listed below) found that input parameters are often ambiguous, especially in retrospective calculations, resulting in less confidence in the results. The software packages which include power analysis provide more confidence, but only when the procedures for calculation are thoroughly explained.

Georgetown University:

- http://members.aol.com/johnp71/postpowr.html;

UCLA:

- http://www.stat.ucla.edu/calculators/powercalc;

EPA beta version:

- http://www.epa.gov/earth1r6/6wq/ecopro/watershd/monitrng/qappsprt/sampling.htm)

Power should be a consideration for any hypothesis test, yet the difficulty in calculating power for nonparametric tests means that it is often ignored. For demonstration purposes, power was considered for the two-sample t-test analyses found in Chapter IV. The powers of these tests were approximated using Minitab™, which has a power analysis calculation for a two-sample t-test (but does not provide an explanation of calculation procedures).

Using the inputs of sample size, minimum detectable difference (chosen to be 10% of the mean of each upstream/background data set), and the standard deviation

(background sigma used as an estimate of an overall sigma), the power of each t-test was actually very low (see results below).

Table V.1: Power Analysis Example

Site	Sample Size	Detectable Difference	Sigma	Power
HM4 BOD5	60	0.1 (10% of 1st 5-yr mean)	0.4	0.2741
HM6 NO3	60	40 (10% of 1st 5-yr mean)	265	0.1299
RO2 NH4	60	5 (10 % of 1st 5-yr mean)	17	0.3588
RO1 RO2 NH4	120	0.3 (10% of upstream mean)	2	0.2120

This means that the t-tests performed in Chapter IV actually had a small chance (all less than 50%) of actually detecting the prescribed difference in means, even with these large sample sizes. Of course, the power increases as the minimum detectable difference required increases (e.g. the power of detecting a difference of 1.0 between sites RO1 and RO2 equals 0.9989), but perhaps not enough to satisfy management concerned with detecting real differences and impacts in the environment. Power also changes as the estimate of standard deviation (sigma) changes. Also, the actual data for these sites did not have the exact sample size included in the power calculation (See Appendix D), this was just an expected value determined by the sampling protocol. In the real world, data can be missing from the record, deceasing the power of the analysis used to detect differences. Power analysis needs a great deal of attention, as it can potentially provide a type of quality control for analysis methods.

Graphical Depiction of Data

One of the simplest ways to help in the interpretation of data is to include a graphical depiction. For example, using time series plots, Q-Q plots, histograms, and box

plots, the analyst can visually interpret the data and make decisions about distribution assumptions, trends, and standards violations with a glance. Plotting the data before analysis was strongly recommended by Montgomery and Reckhow (1984), but in the context of exploratory data analysis. Graphical depictions can aid in the comparability of information from monitoring data by allowing others to 'see' the data, and judge for themselves whether a trend is apparent, or samples have exceeded a standard limit. This type of information should only accompany more scientifically defensible analysis methods, as graphs and pictures can be manipulated with scale, color, or resolution to achieve a desired affect. Careful attention must be paid to the attributes of graphs and pictures, as the analyst can chose a scale to bias the graphical representation, and thus the information conveyed.

Estimation and Confidence Intervals

Another analysis tool, which can be combined with graphics or significance testing, is that of estimates and confidence intervals. "Ordinary confidence intervals provide more information than do p-values. Knowing that a 95% confidence interval includes zero tells one that, if a test of the hypothesis that the parameter equals zero is conducted, the resulting p-value will be greater than 0.05" (Johnson, 1999). A confidence interval gives an estimate of the effect size, as well as a measure of uncertainty, i.e. a confidence interval of (-5, 300) is less well estimated [and potentially embarrassing result] than a parameter with an interval of (120, 130) (Johnson, 1999). Providing a trendline on a time-series plot, with a confidence interval of the slope of that line, can answer questions about trends and compliance without statistical testing.

Meta-Analysis

A type of statistical analysis, called meta-analysis, has been used in the medical and behavioral sciences to combine results from different studies to help draw conclusions about the overall status of the area of interest. This type of analysis might be very useful in the water quality field to combine results from separate studies into one large "picture" of the water quality of a specific river or watershed. Two studies reviewed in Chapter III performed such an analysis (Stoddard, 1998; Brown, 1998). Unfortunately, most water quality studies poorly document the statistical assumptions and parameters that are vital to a meta-analysis study. This restricts the use of such studies in the water quality field. Another perspective is that meta-analysis can reduce dependence on significance testing by examining replicated studies. However, meta-analysis can be dangerously misleading if nonsignificant results, or results that did not conform to the conventional wisdom, were less likely to have been published. (Johnson, 1999)

Interval Testing

Another type of data analysis that shows promise in the water quality field was that described by McBride (1998). Interval testing, though a type of significance testing, allows for the connection between what is statistically significant, and what is ecologically significant. Using the inequivalence hypothesis described by McBride (1993, 1998, 1999a) takes into account extra precaution towards the environment, as the null assumes that an environmental impact has already taken place, and the analysis must

prove that it hasn't. Whereas in any point null hypothesis or equivalence hypothesis, the assumption is that there has been no impact, and the test must prove differently.

Decision Theory

One approach to data analysis is especially related to management is to use statistical decision theory: the theory of acting rationally with respect to anticipated gains and losses, in the face of uncertainty (Johnson, 1999). For example, in most hypothesis testing, the Type I error (rejecting a true null hypothesis) is strictly set at 0.05, yet the type II error (accepting a false null hypothesis) is not examined. Environmentally, a type II error may be more costly, and thus should be taken into account. There are other parameters of water quality (i.e. central tendency, constituent variance or variability in the "natural" environment, biological conditions) that could also be taken into account before a decision is made. This is not unlike the evidentiary or risk assessment process described at the beginning of this chapter.

Biological Assessments

Probably the greatest argument against significance testing is that results may not be biologically or ecologically relevant. "It is not enough to detect differences in lieu of determining an impact's magnitude and cause or in lieu of understanding its consequences. It would be wiser to decide first what is biologically relevant and then use hypothesis testing to look for biologically relevant effects, not merely run a general 'search for significance'." (Karr and Chu, 1998) "Overreliance on statistical correlation, t-tests, or other statistical models can short-circuit the process of looking at data and

asking whether they make sense and what they show. Dependence on p-values can divert scientists and managers from exploring the biology responsible for patterns in data, no matter when or by whom they were collected." (Karr and Chu, 1998)

To better connect monitoring with information about the biological integrity of the waterbody, the EPA has recommended to all states the use of its Rapid Bioassessment Protocol (RBP) modified habitat assessment. The framework of bioassessment consists of characterizing reference conditions upon which comparisons can be made, and identifying appropriate biological attributes with which to measure the condition. These reference conditions are representative of biological health. (Gerritsen and Leppo, 1998). The biological attributes to be measured represent elements of the structure of the ecosystem and are called metrics. A metric is defined as a characteristic of the biota that changes in some predictable way with increased human influence. (Gerritsen and Leppo, 1998)

Sampling of the biological metrics, and assessing the subsequent water quality using a biological index (ranking and scoring) procedure, is becoming increasingly popular in the water quality field. This type of static analysis does not give information about changing conditions (i.e. trends and differences in populations), but can be combined with significance testing to bring real meaning to the monitoring data, both chemical and biological. "The objective of biological monitoring is to detect human-caused deviations from baseline biological integrity, and to evaluate the biological – not statistical – significance of those deviations and their consequences." (Karr and Chu, 1998) "When a study is based on tested biological metrics, hypothesis testing can be appropriate. By providing a biological yardstick for ranking sites according to their

condition, multimetric indexes can answer these questions. Because their statistical properties are known and their statistical power can be calculated, multimetric indexes can be used to compare sites statistically". (Karr and Chu, 1998) Although the same statistical arguments apply to using this type of data for analysis, using biological assessment data in combination with chemical data and appropriate statistical analyses can provide more thorough information about the dynamic condition of the water.

Bayesian Methods

A final statistical analysis approach, which was mentioned briefly in McBride (1998), is that of using a different branch of statistics, called Bayesian statistics. "Bayes' theorem offers a formula for converting between the probability of observed or more extreme data given that the null hypothesis is true (p-value) and the probability that the null hypothesis is true, given the data [for one-sided tests *only*]" (often the information sought in the first place!) (Johnson, 1999).

Bayes' Theorem: $Pr[Ho|data] = Pr[data|Ho] * Pr[Ho] / Pr[data]$

Another, more lucid explanation of this theorem is provided by Carver (1978). "What is the probability of obtaining a dead person (D) given that the person was hanged (H); that is, in symbol form, what is p(D|H)? Obviously, it will be very high, perhaps .97 or higher. Now, let us reverse the question: What is the probability that a person has been hanged (H) given that the person is dead (D); that is, what is p(H|D)? This time the probability will undoubtedly be very low, perhaps .01 or lower. No one would be likely to make the mistake of substituting the first estimate (.97) for the second (.01); that is, to accept .97 as the probability that a person has been hanged given that the person is dead.

Even though this seems to be an unlikely mistake, it is exactly the kind of mistake that is made with the interpretation of statistical significance testing---by analogy, calculated estimates of p(H|D) are interpreted as if they were estimates of p(D|H), when they are clearly not the same." (Carver, 1978)

Using Bayesian approaches, the Pr[Ho], probability of a true null hypothesis, is determined before data are gathered and referred to as the prior probability of Ho. Standard (sometimes referred to as 'frequentist') significance testing considers this probability to be unknown. This prior probability of Ho can be determined subjectively or through objective means. Then, collection of data can update or modify the belief in its value. (Johnson, 1999)

A Bayesian confidence interval (say for 95%) is interpreted to mean that the probability that the true value of the parameter lies in the interval is 95%, as opposed to a standard (frequentist) confidence interval (say for 95%), which interprets to mean that if the study were repeated a large number of times, 95% of the confidence intervals that resulted would contain the true value of the parameter. McBride (2000) Therefore, the Bayesian approach only considers the data obtained, not data that might be obtained if the study were repeated infinitely, nor the data more extreme than that obtained. "For decision analysis, Bayes' theorem offers a very logical way to make decisions in the face of uncertainty. It allows for incorporating beliefs, data, and the gains or losses expected from possible consequences of decisions." (Johnson, 1999) Type I and II errors and p-values are therefore meaningless and not needed.

Comparable Information in Other Fields of Data Collection

One excellent example of the goal for the water quality field is the area of weather reporting. Atmospheric scientists have developed, from a large list of variables and processes, a graphical interpretation of weather conditions that conveys instantly to the user the current state of the weather, what has occurred in the past, and what is likely to happen in the future. The importance of weather in our immediate lives has perhaps been the impetus to create consensus in atmospheric condition assessment. These weather interpretations are transparent, comparable and auditable, as they are standardized and accepted to convey the best information upon which to act.

Another example is the area of economic reporting. Several different indicators and indexes have been developed to aid in interpretation of the daily/monthly/yearly flux of the economy. Graphics, in the form of time series plots of these indexes, are used to convey understanding of trends in various sectors of the economy (Ward, 1998). For example, the Dow Jones Index has become an accepted 'standard' method for reporting a type of economic information upon which management and business decisions are based.

"The indicators and indices have been developed through well-documented and reviewed protocols. This is not to say that there are not disagreements over how the indices are computed, but it does reflect these debates occurring away from day-to-day reporting of the information" (Ward, 1998). "In other words, the science that underpins economic reporting is well developed and documented in protocols that are established on their scientific merit and not their particular outcome" (Ward, 1998).

Conclusions

The above section has outlined just a few of the analysis alternatives that can either replace, or supplement statistical data analysis methods. However, the entrenchment of significance testing in the scientific world, combined with the plethora of analysis alternatives, make it difficult for data analysts to produce comparable information from water quality data analysis.

The subject of this discussion has focused on developing 'standard' guidance for data analysis methods, and how some methods might improve the comparability of information from monitoring. It is obvious that there are many 'right' methods for analysis, yet management is often missing comparable information for decision-making. Management needs information that is dependable, concise, comparable and bias-free in order to make fair and auditable decisions regarding the environment. Arguments about the process through which the information underlying management decision-making was created can only be eliminated through acquisition of comparable information in a manner that is transparent and auditable. Does this call for the development of 'standard' analysis methods?

Development of 'standard' protocols for water quality data analysis is suggested as a means to help this field mature to the same point of confidence about information for management decision-making as observed in weather and economic reporting. This, in turn, could perhaps bring the water quality field closer to the public, allowing water quality monitoring information to be broadly examined, and increasing public support for monitoring efforts.

Chapter VI. Summary, Conclusions and Recommendations

Summary

The previous five chapters of this thesis have fulfilled the tasks outlined in Chapter I: (1) to examine the data analysis methods that are currently being used to analyze water quality monitoring data, as well as the criticisms of using those types of methods; (2) to explore how the selection of methods to analyze water quality data can impact the comparability of information used for water quality management purposes, and; (3) to offer options by which data analysis methods employed in water quality management can be made more transparent and auditable.

These tasks were accomplished through a literature review of criticisms of current data analysis methods (Chapter II), as well as texts, guidance and journals dealing with water quality assessments (Chapter III). Then, the common statistical analysis methods found were applied to the New Zealand Water Quality River Network data set. The purpose being to establish how information changes as analysis methods change, and to determine if the information produced from different data analysis methods was comparable (Chapter IV). The results of the literature review and data analysis were then discussed, highlighting problems with the prevalent use of significance testing in the water quality field. Chapter V further discussed options through which to begin solving these problems and produce comparable information for water quality management decision-making.

Conclusions

For several years it has been known, or suspected, that current methods for producing information from water quality data are subject to misuse and inappropriate application. Lack of statistical knowledge has caused poorly planned method selection and results that are not always comparable. This thesis has documented the problems associated with data analysis method selection for water quality monitoring, in an effort to provide problem definition as the first step in creating a solution. The process of documenting these problems has led to the conclusions discussed below:

1) Reviewing literature on water quality monitoring reveals the frequent use of a common class of statistical procedures (e.g. hypothesis testing) to produce information about water quality from the raw data. The majority of reviewed analysis methods use the concept of "statistical significance" to validate the information produced, be it comparison of means/medians (e.g. upstream/downstream averages), or evaluation of trends, or detection of extremes. It is with these methods that most of our knowledge about the water quality of our nation has been derived. From government monitoring projects to private monitoring studies, it appears from the literature review (Chapter III) that despite recent efforts to provide auditable information, data analysis procedures are often loosely planned and documented and statistical results rarely explained. Except for a few studies of water quality statistics (Harcum et al., 1992; Hirsch, 1988; Montgomery and Reckhow, 1984, Montgomery and Loftis, 1987; Loftis et al., 1989; McBride, 1998, 1999a), alternative analysis methods with which to compare results are never explored,

significance rarely explained, and information, once produced, never questioned, just reported as is. Of course discussions that led up to publication, if they questioned the methods, are rarely shared with the reader.

2) Through EPA's requirements for State 305(b) reports and 303(d) listing of impaired waters, it is apparent that the vision is being developed to create monitoring systems that will produce information that will answer basic questions about our nation's water quality. But when reviewing state assessment methodologies and other water quality studies, it is evident that the analysis procedures fall short of providing indisputable information due to the fact that the assessments are often based on subjective narrative criteria or relatively small monitoring data sets, and lack broadly peer-reviewed and agreed upon data analysis methods.

3) Although the methods selected to produce water quality information are being used correctly, they may not be universally accepted, or appropriate for the type of information about the environment that is needed. The availability of numerous analysis procedures means that methods selected to produce the same type of information (i.e. trends) may be different, resulting in a non-comparable basis for the same management decisions (Chapter IV).

4) Because significance testing methods have been available and accepted for years, their appropriateness has been rarely questioned in the field of water quality monitoring, until now. An argument that is at the forefront of the medical sciences is whether to use

significance testing at all (Chow and Liu, 1992; Loftus, 1991; Royall, 1992; Berger and Berry, 1988). The value of these discussions in medicine is that they illustrate to other scientific fields that there are concerns with creating valid information using hypothesis testing methods for data analysis (McBride, 1993, 1998, 1999a; Johnson, 1999).

5) The solution to producing more valid information for management decision-making depends on the appropriateness of the methods chosen for the type of questions being asked, and the comparability of these methods with other, similar assessments. Many of the supplemental and alternative methods to significance testing discussed in the previous chapters could be utilized to aid in the interpretation of monitoring data, data which is influenced by so many unknown variables that interpretation is often difficult.
The use of new methods that are more appropriate in creating scientifically defensible information is becoming more common in the medical field (Chow and Liu, 1992). However, these methods have not managed to effectively infiltrate water quality monitoring. Medical and epidemiological studies have shown that the use of methods such as meta-analyses, Bayesian statistics, and equivalence testing can produce more objective and valid information from the data than standard significance testing. These alternatives, as well as others, need to be explored for applicability to water quality data analysis, in an effort to produce more comparable information from monitoring.

6) Solutions to the problems documented in this research may not come through common analysis methods, but instead require a deeper understanding of statistical theory, closer connections to the use of the information (i.e. management input), as well

as new thinking about data analysis procedures. These considerations in the development of 'standard' water quality data analysis protocols will help to ensure that the procedures are transparent and auditable, and that results are comparable.

Recommendations

The following recommendations are suggested to help further the endeavor of providing better data analysis methods through which to produce information for management decision-making. These suggestions could be fulfilled through further academic study, interagency cooperative efforts (e.g. state and national water quality monitoring councils), or through a single entity taking the lead in providing guidance for water quality data analysis.

1) The subjects explored in this thesis established that there are many methods available for analysis and interpretation of water quality data. Not only are there statistical methods, but graphical, estimation, Bayesian, and biological methods, to name a few. It was beyond the scope of this thesis to explore the applicability of these methods to water quality data and compare the results with those from hypothesis testing, but such an examination could prove very useful.

2) If hypothesis testing is to continue to be the main venue through which water quality data are interpreted, better attention must be paid to distribution assumptions, flow-adjustment, and power analysis. The first two are easily handled, but the third, power analysis, is a complex subject. Power can be used to determine effective sample sizes to

detect a significant difference fairly easily. However, calculation of the power of certain tests given a sample size can be complicated for parametric statistics, and even more so for nonparametric. Power analysis tools (software, internet calculators) can aid greatly, but a broad review of these tools for comparability of results must first take place in order to ensure quality of results.

3) The recent development of protocols for biological monitoring and assessment methodologies could prove to be the most informative way to assess water quality. These methods are relatively new, and so have not been scrutinized like the methods used to interpret chemical data. Many of the same statistical issues discussed in this thesis apply to biological data as well. The movement towards establishing broadly peer-reviewed methods for data analysis is impending, and all avenues of analysis methods should be thoroughly explored.

The bottom line is that the application of science, individually administered, is not going to make data analysis any easier, or results more comparable. There are too many variables involved, and too many methods through which to explore data. Nevertheless, if management requires accepted, scientifically defensible methods that produce comparable results upon which to base their decisions, consensus must be obtained about what those methods should be. Several documents have been developed for standard methods for sampling protocols and laboratory analysis. Following this trend, it seems only natural to develop standard methods of data analysis as well. As discussed in the Scope section of Chapter I, this should only include methods used for management

decision-making. Exploratory data analysis employed by researchers needs to remain untethered and flexible.

This is an issue that can only partially be resolved through science. Research, such as this thesis, can establish that there are common methods being used, compare the results obtained with differing methods, and document that there are problems with current data analysis procedures. But the decision-makers who are knowledgeable about monitoring resources, costs, and consequences of individual decisions will need to be the ones who, through a fair and open process, develop a guidance of acceptable methods for water quality monitoring data analysis.

List of References

Abeyta, C.G. and R.G. Roybal. 1996. Ground-water quality, water year 1995, and statistical analysis of ground-water quality data, water years 1994-95, at the chromic acid pit site, U.S. Army Air Defense Artillery Center and Fort Bliss, El Paso, Texas, *U.S. Geological Survey Water-Resources Investigations Report 96-4211*, U.S. Geological Survey.

Adkins, N.C. 1992. A framework for development of data analysis protocols for groundwater quality monitoring systems. Ph.D. Thesis, Colorado State University.

Arizona DEQ. 2000. Assessment criteria: draft for 2000 water quality assessment report. Arizona Department of Environmental Quality.

Arthur, M.A., Coltharp, G.B. and D.L. Brown. 1998. Effects of best management practices on forest streamwater quality in Eastern Kentucky. *Journal of the American Water Resources Association, Paper No. 97106* 34(3):481-495.

Baldys, S., L.K. Ham and K.D. Fossum. 1995. Summary statistics and trend analysis of water quality data at sites in the Gila River Basin, New Mexico and Arizona, *U.S. Geological Survey Water-Resources Investigations Report 95-4083*, U.S. Geological Survey.

Barath, M. 2000. EPA Region III 305(b) Coordinator, U.S. Environmental Protection Agency. Personal Communication.

Becker, D.S. and J.W. Armstrong. 1988. Development of regionally standardized protocols for marine environmental studies. *Marine Pollution Bulletin* 19(7):310-313.

Berger, J.O. and D.A. Berry. 1988. Statistical analysis and the illusion of objectivity. *American Scientist* 76:159-165.

Berkson, J. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association* 33:526-542.

Berndt, M.P. 1996. Ground-water quality assessment of the Georgia-Florida Coastal Plain study unit – analysis of available information on nutrients, 1972-92, *U.S. Geological Survey Water Resources Investigations Report 95-4039*, U.S. Geological Survey, Tallahassee, Florida.

Bexfield, L.M. and S.K. Anderholm. 1997. Water-quality assessment of the Rio Grande Valley, Colorado, New Mexico, and Texas – ground-water quality in the Rio Grande Flood Plain, Cochiti Lake, New Mexico, to El Paso, Texas, 1995, *USGS Water Resources Investigations Report 96-4249*, U.S. Geological Survey, Albuquerque, New Mexico.

Bollinger, S.W. and D.L. Sitlinger. 1997. Water quality of interstate streams in the Susquehanna River Basin. *Susquehanna River Basin Commission Publication 185.*

Brown, D.W., McCain, B.B. Horness, B.H. Sloan, C.A., Tilbury K.L., Pierce S.M., Burrows, D.G., Chan, Sin-Lam, Landahl, J.T. and M.M. Krahn. 1999. Status, correlations, and temporal trends of chemical contaminants in fish and sediment from selected sites on the Pacific Coast of the U.S.A. *Marine Pollution Bulletin* 37(1-2):67-85.

Bryers, G.G. 1999. National Rivers Water Quality Network: Data and Meta-Data. National Institute of Water and Atmospheric Research, Hamilton, New Zealand.

Butler, D. L. 1996. Trend analysis of selected water quality data associated with salinity control projects in the Grand Valley, Lower Gunnison River Basin, and at Meeker Dome, Western Colorado, *U.S. Geological Survey Water-Resources Investigations Report 95-4274*, U.S. Geological Survey.

Carver, R.P. 1978. The case against statistical testing. *Harvard Educational Review* 48:378-399.

Chatfield, C. 1985. The initial examination of data. *Journal of the Royal Statistical Society*, Series A 148:214-253.

Chow, S.C. and J.P. Liu. 1992. *Design and Analysis of Bioavailability and Bioequivalence Studies.* Marcel Dekker, New York.

Clifton, D. 1985. Analysis of biological data collected in the Bull Run Watershed, Portland, Oregon, 1978 to 1983, *U.S. Geological Survey Water Resources Investigations Report 85-4245*, U.S. Geological Survey.

Clow, D.W. and M.A. Mast. 1999. Long-term trends in stream water and precipitation chemistry at five headwater basins in the Northeastern United States. *Water Resources Research* 35(2):541-554.

Colman, J.A. and S.F. Clark. 1994. Geochemical data on concentrations of inorganic constituents and polychlorinated biphenyl congeners in streambed sediments in tributaries to Lake Champlain in New York, Vermont, and Quebec, 1992, *U.S. Geological Survey Open-File Report 94-472*, U.S. Geological Survey.

Copeland, R. 2000. Florida Department of Environmental Protection. Personal Communication.

Deacon, J.R. and N.E. Driver. 1999. Distribution of trace elements in streambed sediment associated with mining activities in the Upper Colorado River Basin, Colorado, USA, 1995-96. *Archives of Environmental Contamination and Toxicology* 37:7-18.

Dennehy, K.F. Litke, D.W., McMahon, P.B., Heiny, J.S. and C.M. Tate. 1995. Water quality assessment of the South Platte River Basin, Colorado, Nebraska, and Wyoming – analysis of available nutrient, suspended-sediment, and pesticide data, water years 1980-92, *U.S. Geological Survey Water-Resources Investigations Report 94-4095*, U.S. Geological Survey.

Denton, G. 2000. Tennessee Water Pollution Control Division. Personal Communication.

DuPont, W.D. 1983. Sequential stopping rules and sequentially adjusted p-vaues: does one require the other? *Controlled Clinical Trials* 4(1):3-10.

Edwards, R.E. 1998. The 1998 Susquehanna River Basin water quality assessment 305(b) report. *Susquehanna River Basin Commission Publication 201.*

EPA. 1989. 1992 addendum. Statistical analysis of ground-water monitoring data at RCRA facilities: interim final guidance. U. S. Environmental Protection Agency Office of Solid Waste, Washington D.C.

EPA. 1992. Monitoring guidance for the national estuary program, *EPA-842-B-92-004.* U. S. Environmental Protection Agency Office of Water, Washington D.C.

EPA. 1997a. Guidelines for preparation of the comprehensive state water quality assessments (305 (b) Reports) and electronic updates, *EPA-841-B-97-002A.* U.S. Environmental Protection Agency Office of Water, Washington D.C.

EPA. 1997b. Information collection rule: draft data analysis plan. U.S. Environmental Protection Agency Office of Water, Washington D.C.

EPA. 1997c. Monitoring guidance for determining the effectiveness of nonpoint source controls, *EPA-841-B-96-004.* U.S. Environmental Protection Agency Office of Water, Washington D.C.

EPA. 1998. Report of the federal advisory committee on the total maximum daily load (TMDL) program, *EPA 100-R-98-006*. The National Advisory Council for Environmental Policy and Technology, U. S. Environmental Protection Agency Office of the Administrator.

EPA. 1999. Technical guidance on monitoring and data interpretation to support implementation of water quality standards, assessment of water quality standards attainment under Section 305(b), and listing and delisting of threatened and impaired waters under Section 303(d): draft outline. U.S. Environmental Protection Agency.

Fleiss, J.L. 1987. Some thoughts on two-tailed tests. *Controlled Clinical Trials* 8:394.

GAO. 2000. Water quality: key EPA and state decisions limited by inconsistent and incomplete data, *GAO/RCED-00-54*. U.S. General Accounting Office.

Gerritsen, J. and E.W. Leppo. 1998. Development and testing of a biological index for warmwater streams of Arizona. Prepared for Arizona Department of Environmental Quality by Tetra Tech, Inc.

Gilbert, R.O. 1987. *Statistical methods for environmental pollution monitoring*. Van Nostrand Reinhold, New York.

Good, I.J. 1982. Standardized tail-area probabilities. *Journal of Statistical Computation and Simulation* 16:65-66.

Goodman, S. 1988. One-sided or two-sided p-values? *Controlled Clinical Trials* 9:387-388.

Goodman, S. N. 1993. P-values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology* 137(5):485-496.

Harcum, J.B., J.C. Loftis, and R.C. Ward. 1992. Selecting trend tests for water quality series with serial correlation and missing values. *Water Resources Bulletin* 28(3):469-478.

Havens, K.E., Flaig, E.G., James, R.T., Lostal, S. and D. Muszick. 1996. Results of a program to control phosphorus from dairy operations in South-Central Florida, USA. *Environmental Management* 20(4):585-593.

Heiskary, S., Lindbloom, J. and C.B. Wilson. 1994. Detecting water quality trends with citizen volunteer data, Minnesota Pollution Control Agency. *Lake and Reservoir Management* 9(1):4-9.

Helsel, D.R. and R.M. Hirsch. 1992. *Statistical methods in water resources: studies in environmental science 49.* U.S. Geological Survey Water Resources Division, Reston, Virginia. Elvesier, New York.

Intelligent Decision Technologies, Ltd. 1998. *WQStat Plus™ User's Guide.*

Johnson, D.H. 1995. Statistical sirens: the allure of nonparametrics. *Ecology* 76:1998-2000.

Johnson, D.H. 1999. The insignificance of statistical significance testing. *Journal of Wildlife Management* 63(3):763-772.

Karr, J.R. and E.W. Chu. 1998. *Restoring life in running waters: better biological monitoring.* Island Publishing.

Kennedy, K. 1995. A statistical analysis of TxDOT highway storm water runoff: comparisons with the existing North Central Texas municipal storm water database. North Central Texas Council of Governments.

Kirkland, G. 2000. South Carolina 305(b) Coordinator. Personal Communication.

Koebel, J.W. Jr., Jones, B.L. and D.A. Arrington. 1999. Restoration of the Kissimmee River, Florida: water quality impacts from canal backfilling. *Environmental Monitoring and Assessment* 57:85-107.

Kress, N., Hornung, H. and B. Herut. 1998. Concentrations of Hg, Cd, Cu, Zn, Fe and Mn in deep sea benthic fauna from the southeastern Mediterranean Sea: a comparison study between fauna collected at a pristine area and at two waste disposal sites. *Marine Pollution Bulletin* 36(11):911-921.

Lapp, P., Madramootoo, C.A., Enright, P., Papineau, F. and J. Perrone. 1998. Water quality of an intensive agricultural watershed in Quebec. *Journal of the American Water Resources Association, Paper No. 97033* 34(2):427-437.

Lavenstein, G.G. and K.D. Daskalakis. 1998. U.S. long-term coastal contaminant temporal trends determined from mollusk monitoring programs, 1965-1993. *Marine Pollution Bulletin* 37(1-2):6-13.

Loftis, J.C. 1989. An evaluation of trend detection techniques for use in water quality monitoring programs, *EPA-600-3-89-037.* U.S. Environmental Protection Agency Office of Research and Development, Environmental Research Laboratory, Corvallis, Oregon.

Loftis, J.C., Iyer, H.K. and H.J. Baker. 1999. Rethinking Poisson-based statistics for groundwater quality monitoring. *Groundwater* 37(2):275-280.

Loftus, G.R. 1991. On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology* 36:102-105.

MacDonald, L.H. 1994. Developing a monitoring project. *Journal of Soil and Water Conservation* May-June:221-227.

Marsh, D. 2000. Assessment Coordinator, Arizona Department of Environmental Quality. Personal Communication.

Martin, J.D. and C.G. Crawford. 1987. Statistical analysis of surface-water quality data in and near the mining region of southwestern Indiana, 1957-80, *U.S. Geological Survey Water-Supply Paper 2291*, U.S. Geological Survey.

Matthews, R. 1998. *Lying science.* Sunday Star-Times, Auckland, New Zealand. September 27.

Mattraw, H.C., D.J. Scheidt and A.C. Federico. 1987. Analysis of trends in water quality data for water conservation area 3A, the Everglades, Florida, *U.S. Geological Survey Water-Resources Investigations Report 87-4142*, U.S. Geological Survey.

McBride, G.B., Loftis, J.C. and N.C. Adkins. 1993. What do significance tests really tell us about the environment? *Environmental Management* 17(4):423-432.

McBride, G.B. 1998. Statistical methods: when differences are equivalent. *Water & Atmosphere* 6(4):21-23. National Institute of Water and Atmospheric Research, New Zealand.

McBride, G.B. 1999a. Equivalence tests can enhance environmental science and management. *Australia & New Zealand Journal of Statistics* 41(1):19-29.

McBride, G. B. 1999b. National Institute of Water and Atmospheric Research, Hamilton, New Zealand. Provision of Interval Test Algorithm in MS-Excel™. Personal Communication.

McBride, G.B. 1999c. National Institute of Water and Atmospheric Research, Hamilton, New Zealand. Thesis consultation. Personal Communication.

McBride, G.B. 2000. Some issues in statistical inference: draft. National Institute of Water and Atmospheric Research, Hamilton, New Zealand.

McGill, R., Tukey, J.W. and W.A. Larsen. 1978. Variations of box plots. *The American Statistician* 32(1).

McMahon, G. and D.A. Harned. 1998. Effect of environmental setting on sediment, nitrogen, and phosphorus concentration in Albemarle-Pamlico Drainage Basin, North Carolina and Virginia, USA. *Environmental Management* 22(6):887-903.

Mintab, Inc. 1997. Minitab™ Version 12.

Momen, B., Eichler, L.W., Boylen, C.W. and J.P. Zehr. 1997. Are recent watershed disturbances associated with temporal and spatial changes in water quality of Lake George, New York USA? *Environmental Management* 21(5):725-732.

Montgomery, R.H. and J.C. Loftis. 1987. Applicability of the t-test for detecting trends in water quality variables. *Water Resources Bulletin, Paper No. 86130*, 23(4).

Montgomery, R.H. and K.H. Reckhow. 1984. Techniques for detecting trends in lake water quality. *Water Resources Bulletin, Paper No. 82105* 20(1):43-52.

Mueller, D.K. 1990. Analysis of water quality data and sampling programs at selected sites in North-Central Colorado, *U.S. Geological Survey Water-Resources Investigations Report 90-4005,* U.S. Geological Survey.

Mueller, D.K. 1995. Nutrients in ground water and surface water of the United States: an analysis of data through 1992, *U.S. Geological Survey Water-Resources Investigations Report 95-4031*, U.S. Geological Survey.

NJ Department of Environmental Protection. 1999. New Jersey water quality inventory report 1998, *Environmental Planning and Science, Division of Science, Research, and Technology Water Assessment Team File Report.* New Jersey Department of Environmental Protection.

NYS Department of Environmental Conservation. 1998. New York State water quality, 1998. Bureau of Watershed Assessment and Research, Division of Water, New York State Department of Environmental Conservation.

Nester, M.R. 1996. An applied statistician's creed. *Applied Statistics* 45:401-410.

Nimmo, D.R., Willox, M.J., LaFrancois, T.D., Chapman, P.L., Brinkman, S.F. and J.C. Greene. 1998. Effects of metal mining and milling on boundary waters of Yellowstone National Park, USA. *Environmental Management* 22(6):913-926.

North Caroline State University Cooperative Extension. 1999. Detecting water quality changes before and after BMP implementation: use of SAS for statistical analysis, *NWQEP Notes: The NCSU Water Quality Group Newsletter Number 93.*

Nunnally, J.C. 1960. The place of statistics in psychology. *Educational and Psychological Measurement* 20:641-650.

Oklahoma Water Resources Board. 1999. Subchapter 15: use support assessment protocols: draft. Water Quality Programs Division, Oklahoma Water Resources Board.

Parahos, R., Pereira, A.P. and L.M. Mayr. 1998. Diel variability of water quality in a tropical polluted bay. *Environmental Monitoring and Assessment* 50(2):131-141.

PEER: Public Employees for Environmental Responsibility. *Murky Waters: Official Water Quality Reports are All Wet, An Inside Look at EPA's Implementation of the Clean Water Act*. May 1999. http://www.peer.org/execsum.html

Pinsky, P., Lorber, M., Johnson, K., Kross, B., Burmeister, L., Wilkins, A. and G. Hallberg. 1997. A study of the temporal variability of atrazine in private well water, part II: analysis of data. *Environmental Monitoring and Assessment* 47(2):197-221.

Preece, D.A. 1990. R.A. Fisher and experimental design: a review. *Biometrics* 46:925-935.

Redfield, G. ed. 1999. Everglades interim report. South Florida Water Management District.

Reif, M. 2000. Alabama Water Quality Report to Congress Coordinator, Alabama Department of Environmental Management, Water Division: Water Quality Section. Personal Communication.

Richard, N. 2000. California State 305(b) Coordinator. Personal Communication.

Rinella, J.F. 1986. Analysis of fixed-station water quality data in the Umpqua River Basin, Oregon, *U.S. Geological Survey Water Resources Investigations Report 85-4253*, U.S. Geological Survey.

Royall, R.M. (eds. J.M. Bernardo, J.O. Berger, A.P. David and A. Smith) 1992. The elusive concept of statistical evidence. *Bayesian Statistics* 4:405-418. Oxford University Press, New York.

Sample, L.J., Steichen, J. and K.R. Kelley Jr. 1998. Water quality impacts from low water fords on military training lands. *Journal of the American Water Resources Association, Paper No. 96165* 34(4):939-949.

Santillo, D., Stringe, R.L., Johnston, P.A. and J. Tickner. 1998. The precautionary principle: protecting against failures of scientific method and risk assessment. *Marine Pollution Bulletin* 36(12):939-950.

Savage, I.R. 1957. Nonparametric statistics. *Journal of the American Statistical Association* 52:331-344.

Shabman, L.A., Hershner, C., Kator, H.I., Smith, E.P., Smock, L.A., Younos, T., Shaw, L.Y and C.E. Zipper. 1998. Report of the Water Quality Academic Advisory Committee. *Virginia Water Resources Research Center Report No. SR8-1998.* http://www.vwrrc.vt.edu/publications/special.htm

Smith, D.G., McBride, G.B., Bryers, G.G, Wisse, J and D.F.J. Mink. 1996. Trends in New Zealand's national river water quality network. *New Zealand Journal of Marine and Freshwater Research* 30:485-500.

Smith, R.A., R.B. Alexander and M.G. Wolman. 1987. Analysis and interpretation of water-quality trends in major U.S. rivers, 1974-81, *USGS Water Supply Paper 2307*. United States Government Printing Office.

Snedecor, G.W. and W.G. Cochran. 1980. *Statistical Methods*. 7th ed. The Iowa State University Press, Ames, Iowa.

Snyder, N.J., Mostaghimi, S., Berry, D.E., Reneau, R.B., Hong, S., McClellen, P.W. and E.P. Smith. 1998. Impact of riparian forest buffers on agricultural nonpoint source pollution. *Journal of the American Water Resources Association, Paper No. 96132* 34(2):385-395.

Spooner, J., Maas, R.P., Dressing, S.A., Smolen, M.D. and F.J. Humenik. 1985. Appropriate designs for documenting water quality improvements from agricultural NPS control programs, *EPA 440/5-85-001*. U.S Environmental Protection Agency Perspectives on Nonpoint Source Pollution:30-34.

Stoddard, J.L., Driscoll, C.T., Kahl, J.S. and J.H. Kellogg. 1998. A regional analysis of lake acidification trends for the northeastern U.S. 1982-1994. *Environmental Monitoring and Assessment* 51:399-413.

Stoe, T.W. 1998. Water quality and biological assessment of the Wiconisco Creek watershed. *Susquehanna River Basin Commission Publication No. 193*.

Swanek, R. 2000. North Carolina Department of Environment and Natural Resources, Division of Water Quality Planning. Personal Communication.

Takita, C.S. 1998. Nutrients and suspended sediment transported in the Susquehanna River Basin, 1994-96, and loading trends, calendar years 1985-96. *Susquehanna River Basin Commission Publication No.194*.

Teruya, T. 2000. Hawaii State 305(b) Coordinator. Personal Communication.

Vaill, J.E. and D.L. Butler. 1999. Streamflow and dissolved-solids trends, through 1996, in the Colorado River Basin upstream from Lake Powell – Colorado, Utah and Wyoming, *USGS Water Resources Investigations Report 99-4097*, U.S. Geological Survey, Denver, Colorado.

VanArsdall, T. 2000. Kentucky Division of Water. Personal Communication.

Virginia Department of Environmental Quality. 1999. Water quality assessment guidance manual for Y2000: 305(b) water quality report and 303(d) TMDL priority list, revised draft 09/09/99. Virginia Department of Environmental Quality.

Ward, R.C. and J.C. Loftis. 1983. Incorporating the stochastic nature of water quality into management. *Journal of the Water Pollution Control Federation* 55:408-414.

Ward, R.C., J.C. Loftis and G.B. McBride. 1986. The "data-rich but information poor" syndrome in water quality modeling. *Environmental Management* 10(3):291-297.

Ward, R.C. 1998. Management and Monitoring of Water Quality: CB/CE 545 Fall Class Notes. Colorado State University.

Ward, R.C., J.C. Loftis and G.B. McBride. 1990. *Design of Water Quality Monitoring Systems*. Van Nostrand Reinhold, New York.

Wells, F.C. and T.L. Schertz. 1983. Statistical summary of daily values data and trend analysis of dissolved-solids data at national stream quality accounting network (NASQAN) stations, U.*S. Geological Survey Water Resources Investigations Report 83-4172,* U.S. Geological Survey.

Younos, T.M., Mendez, A., Collins, E.R. and B.B. Ross. 1998. Effects of a dairy loafing lot-buffer stream on stream water quality. *Journal of the American Water Resources Association, Paper No. 97040* 34(5):1061-1069.

Appendix A.
Data and Results from McBride (1998)

Appendix A. Data and Results from McBride (1998)

Null and equivalence hypothesis tests and Bayesian probabilities on taxonomic richness data, from Quinn et al. 1992 (Hydrobiologia 248: 235-247). There are seven replicates (in runs) from upstream and from downstream of gold mining operations in six streams.

INPUT DATA ("U" & "D" appellations denote upstream and downstream; "d" denotes difference)

	German Gully		*Houhou*		*Kaniere*		*Kapitea*		*Red Jacks*		*Waimea*	
Replicate	**GU**	**GD**	**HU**	**HD**	**KnU**	**KnD**	**KpU**	**KpD**	**RU**	**RD**	**WU**	**WD**
1	21	8	13	11	8	10	15	8	19	14	13	12
2	16	7	15	10	11	13	21	7	16	14	12	14
3	18	7	16	11	9	8	16	9	18	13	18	12
4	15	12	15	12	15	8	20	10	17	15	11	11
5	14	10	19	11	15	8	17	8	21	14	14	15
6	18	9	14	12	10	8	23	9	18	11	11	13
7	12	8	20	14	11	9	21	9	25	13	13	14

dL%	-20	lower bound of environmentally significant %age change in upstream value
dU%	20	upper bound of environmentally significant %age change in upstream value
alpha	5	maximum permissible probability of rejecting H0 for any comparison, IF that hypothesis is actually true (not that we will ever know for sure)

NB. If the overall significance level is to be controlled (e.g., to 5%), alpha must usually be reduced to a lower value. The most pessimistic reduction is the Bonferroni correction: alpha = 1-(0.95)^1/6 = 0.85% (there being 6 comparisons to be made). I say "usually" because the correction needs to account only for the number of cases where H0 is in fact true. One could argue that it need never be made for the two-sided difference test, because its H0 is never true for observational data like these! And if half the "H0: equivalence" cases were true (and so half were not) the correction would be alpha = 1-(0.95)^1/3 = 1.7%.

RESULTS
SUMMARY

H0: no difference	Sig. diff.	Sig. diff.	No sig. diff.	Sig. diff.	Sig. diff.	No sig. diff.
H0: inequivalence	Inequiv.	Inequiv.	Inequiv.	Inequiv.	Inequiv.	Equiv.
H0: equivalence	Inequiv.	Equiv.	Equiv.	Inequiv.	Equiv.	Equiv.
Bayesian posterior probability (%) that the true difference is within the equivalence interval (using uniform priors)						
	0.33	14.04	53.33	0.01	7.71	97.06

CALCULATED SAMPLE SIZES, DEGREES OF FREEDOM AND CRITICAL t VALUES

Number of replicates, nU = nD = 7	nu = 2(nU + nD -2) = 12	t[alpha(2),nu] = 2.179	t[alpha(1),nu] = 1.782

NB. "alpha(2)" means that we are using the upper AND lower tails of the t-distribution, there being an area = alpha/2 in each. This is used in the two-sided difference tests shown below, and is calculated from Excel's function TINV(alpha,2(n-1)). But "alpha(1)" means that we are considering only the upper tail of the t-distribution, containing an area = alpha. This is used in equivalence tests (which are in effect an amalgam of two one-sided tests). Because the TINV function gives only the two-tailed inverse (i.e., abscissa) of the t-distribution, we must use t[alpha(1),2(n-1)] = TINV(2*alpha,2(n-1)).

Appendix A. McBride (1998)

DERIVED DATA

Median	16	8	15	11	11	8	20	9	18	14	13	13
Means (muU, muD)	16.29	8.71	16.00	11.57	11.29	9.14	19.00	8.57	19.14	13.43	13.14	13.00
SD (standard deviation)	2.98	1.80	2.58	1.27	2.75	1.86	3.00	0.98	3.02	1.27	2.41	1.41
CV (= SD/mu, %)	18.3	20.6	16.1	11.0	24.4	20.4	15.8	11.4	15.8	9.5	18.3	10.9
sp [= sqrt{sum(SD^2)}]		2.46		2.04		2.35		2.23		2.32		1.98
SE [= sp*sqrt(2/n)]		1.32		1.09		1.26		1.19		1.24		1.06
dhat (= muD - muU)		-7.57		-4.43		-2.14		-10.43		-5.71		-0.14
dL (= muU*dL%/100)		-3.26		-3.20		-2.26		-3.80		-3.83		-2.63
dU (= muU*dU%/100)		3.26		3.20		2.26		3.80		3.83		2.63
|100*dhat/muU| (%)		46.5		27.7		19.0		54.9		29.9		1.1
T (= |dhat|/SE)		5.75		4.07		1.71		8.75		4.61		0.14
Ta [= (dhat-dL)/SE]		-3.28		-1.13		0.09		-5.56		-1.52		2.35
Tb [= (dhat-dU)/SE]		-8.22		-7.01		-3.50		-11.93		-7.70		-2.62
F(Ta) (cumulative t,%)		0.33		14.04		53.55		0.01		7.71		98.18
F(Tb) (cumulative t,%)		0.00		0.00		0.22		0.00		0.00		1.11

Appendix B.
Arizona Assessment Criteria

Appendix B1. Arizona Assessment Criteria Using Numeric Standards

DESIGNATED USES AND CONSTITUENTS	NUMBER OF SAMPLES	ASSESSMENT CRITERIA
All uses	Only 1 sample	Cannot assess based only on one water chemistry sample.
Aquatic and Wildlife. -- Toxic Substance	Less than 10 samples (more than 1 sample)	1 sample exceeds = partial support More than 1 sample exceeds = discretion in choosing partial or non-support based on number of samples magnitude of exceedances.
	10 or more samples	**Toxic substances -- Acute criteria** 1 sample exceeds standard = full support 2 or more samples exceed standard = non-support
	4 consecutive days of samples	**Toxic substances -- Chronic criteria** Mean exceeds standard
Aquatic and Wildlife -- Nontoxic substance (except nutrients) and Full Body/Partial Body Contact, Agriculture Irrigation/Livestock Water -- Toxic or Non Toxic Substances	Less than 10 samples (more than 1 sample)	1 sample exceeded standards = partial support. More than 1 sample exceeds standards = partial or non-support based on number of samples and magnitude of exceedances.
	More than 10 samples	Less than 10% samples exceed = full support 10-25% samples exceed = partial support More than 25% samples exceed = non-support
Full Body Contact	Minimum number established in Rules.	**Geometric mean for bacteria** testing during the past two years: Geometric mean repeatedly exceeded = nonsupport Geometric mean exceeded only once = partial support
Nutrients (nitrogen or phosphorus) for Aquatic and Wildlife Uses	More than 1 sample	"Single sample" criteria exceeded Less than 10% samples exceed = full support 10-25% samples exceed = partial support More than 25% samples exceed = nonsupport
	Minimum number established in Rules.	Annual mean standard or 90% standard is exceeded = partial or non-support depends on number of times exceeded in a 5 year period and whether there is substantiating evidence of negative impacts (i.e., fish kills)
Fish Consumption and Domestic Water Source Uses	More than 2 samples	Median of all samples exceeds standard = non-support.
Trends in Water Chemistry Use dependent on parameter.	Sampling periods 10 years apart and > 10 samples per period.	Downward trend, such that standard may be exceeded within the next assessment cycle = full support but "threatened."

(Arizona Department of Environmental Quality Assessment Criteria, 2000)

Appendix B2. Arizona Assessment Criteria Using Narrative Standards

CONSTITUENT AND DESIGNATED USES	NUMBER OF SAMPLES	ASSESSMENT CRITERIA
Fish Consumption		**Fish consumption advisory** = non support
		Off-flavor in aquatic organisms or waterfowl documented.= partial support.
	Used only as supporting evidence	**Fish tissue concentration** median value is above narrative standards assessment guidance = Full (use as weight of evidence and flag for potential problems)
Aquatic and Wildlife	Used only as supporting evidence	**Fish tissue** concentration median value is above narrative standards assessment guidance = flag for potential problems. Contact USFWS, AGFD, or other expert to determine whether "toxic" impacts documented.
		"Narrative toxic standard" -- Impacts to aquatic and wildlife documented (i.e., fish kills or anomalies). (See "toxic" definition in Appendix A.)
	Used only as supporting evidence.	**Index of Biological Integrity (Bioassessments)**: See explanation on page C-4 of this appendix.
Aquatic and Wildlife or Full Body/Partial Body Contact	Used only as supporting evidence.	**Contaminated sediment** median value exceeds criterion
		"Narrative nutrient standard" -- Noxious weeds or algal blooms documented along with elevated pH or low dissolved oxygen. Partial support or non-support based on how often and severe.
		Excessive sedimentation documented = partial support.
Full Body/Partial Body Contact		**Objectionable odor** is documented = partial support. **Water color change** from background levels documented = partial support.
Domestic Water Source		**Drinking water advisory:** Within the past two years related to source water quality of surface water: Advisory issued for less than one week per year = partial support. Advisory issued for more than one week per year = non-support. **Off-taste or odor** in drinking water documented = partial support. **Cause a violation of an aquifer water quality standard** (or contribute to a violation.).= non-support.
Full Body Contact		**Swimming area closures** within the past two years: Less than one week closure per year = partial support. Greater than one week closure per year = non-support.

(Arizona Department of Environmental Quality Assessment Criteria, 2000)

Appendix B3. Arizona Trophic Classification Thresholds

Trophic State	Trophic Status Index	Chlorphyll-*a*	Secchi Depth (meters)	Total Phosphorus (P) (µg/l)		Total Nitrogen (N) (mg/l)	
				Phosphorus-Limited	N& P-Limited	Nitrogen-Limited	N& P-Limited
Oligotrophic	<30	<5	>3	<10	<13	<0.25	<0.28
Mesotrphic	30-45	5-12	1.2-3	10-20	13-35	0.25-0.65	0.28-0.75
Eutorphic	45-65	12-20	0.6-1.2	20-35	35-65	0.65-1.1	0.75-1.2
Hypereutrophic	>65	>20	<0.6	>35	>65	>1.1	>1.2

"Nitrogen-Limited" = N:P ratio is <10
"Phosphorus-Limited" = N:P ratio is >30
"N&P-Limited" = Colimited = N:P ratio is 10-30

Trophic Classification based on: Brezonik, Patrick L., "Trophic State Indices: Rationale for Multivariate Approaches", Lake and Reservoir Management, pp 441-445.

(Arizona Department of Environmental Quality Assessment Criteria, 2000)

Appendix C.
Virginia Designated Use Assessment Criteria

Appendix C1: Virginia's Designated Use Assessment Criteria

	Fully Supporting	Fully Supporting but Threatened	Partially Supporting	Not Supporting
Conventional Pollutants	R ≤10%	NA	11% ≤ R ≤ 25%	R > 25%
Toxic Pollutants	No more than 1 exceedance in a 3 year period (10 sample minimum)	* See fish tissue and sediment criteria	R > 1 Exceedance but ≤ 10% of samples (10 sample minimum)	R > 10 % samples (10 sample minimum)
Biological Data	Not Impaired or Slightly Impaired	Unconfirmed, Moderately Impaired, Evaluated data show potential WQ problems	Confirmed Moderately Impaired or degraded (or two surveys shows moderate impairment)	Severely Impaired or Degraded
Fish Consumption Advisories or Restrictions	None	NA	An advisory from VDH is in place	A restriction from VDH is in place
Shellfish Restrictions or Prohibitions	None	Area classified as **Conditionally Approved** (includes seasonal condemnations)	Areas classified as **Restricted**	Areas classified as **Prohibited** (exception: VPDES outfall areas)
Beach Closures	None	One short term VDH closure with low probability of recurrence (pollution source transient and no VDH plans to implement any control measures)	One or more VDH closure with medium probability of recurrence (VDH preparing plans to implement controls measures)	One or more VDH closure with high probability of recurrence (VDH initiates plans to implement controls measures)
Drinking Water Source Closures	None	One short term VDH closure with low probability of recurrence (pollution source transient and no VDH plans to implement any control measures)	One or more VDH closure with medium probability of recurrence (VDH preparing plans to implement controls measures)	One or more VDH closure with high probability of recurrence (VDH initiates plans to implement controls measures)

*** Fish Consumption Criteria**	* Sediment Criteria
If one or more Level 1 samples exceed one or more risk based SV's – threatened for fish consumption Cause: violation of SV for affected parameter Source: unknown	If one or more ER-M SV(s) or if no ER-M exists, 99th percentile SV exceed – threatened for aquatic life. Cause: violation of SV for affected parameter

R = arithmetic percent violation rate; SV = screening value; ER-M = effects range – medium value
*No water body should be designated impaired (partially or not supporting) based on Level 1 Fish tissue or Sediment or data alone. (Virginia Department of Environmental Quality, 1999)

Appendix C2. Virginia Use Support Assessment

NO.	DESIGNATED USE	SUPPORT OF USE ASSESSMENT CRITERIA
1	Aquatic Life Use	Conventional Pollutants (DO, pH, Temp.); Toxics in water column; Fish tissue and sediments; Biological evaluation.
1a.	Fish Consumption Use	Advisories and restrictions issued by VDH; Comparison of water column data to human health standards; Comparison of fish tissue data to national screening values.
1b.	Shellfish Consumption Use	Restrictive actions for harvesting and marketing of shellfish resources made by Div. Of Shellfish Sanitation of VDH; comparison of data to water quality bacteria standards applicable to designated shellfish waters.
2	Swimming Use	Conventional Pollutant (Fecal Coliform Bacteria) and/or VDH beach closures.
3	Public Water Supply Use	Closures or advisories by VDH; comparison of data to applicable public water supply standards.

(Virginia Department of Environmental Quality, 1999)

Appendix D.
Data and Meta-Data for New Zealand River Network (Bryers, 1999)

Appendix D1. New Zealand Data - Site HM4_BOD5

Date	Site	BOD5 (ppm)	Date	Site	BOD5 (ppm)	Date	Site	BOD5 (ppm)
890125	HM4	1.10	920513	HM4	1.20	950913	HM4	0.80
890222	HM4	0.95	920617	HM4	1.15	951011	HM4	1.40
890323	HM4	1.50	920715	HM4	0.95	951108	HM4	1.65
890419	HM4	1.20	920812	HM4	1.20	951213	HM4	2.25
890524	HM4	1.10	920917	HM4	0.75	960117	HM4	1.10
890614	HM4	1.20	921013	HM4	0.80	960214	HM4	0.80
890712	HM4	1.30	921120	HM4	1.10	960313	HM4	0.85
890817	HM4	0.80	921218	HM4	1.10	960417	HM4	1.00
890913	HM4	0.85	930113	HM4	2.00	960515	HM4	0.75
891018	HM4	1.75	930217	HM4	1.60	960612	HM4	0.90
891116	HM4	2.05	930317	HM4	1.45	960717	HM4	0.60
891214	HM4	1.95	930414	HM4	1.15	960814	HM4	0.85
900117	HM4	1.20	930513	HM4	0.75	960918	HM4	1.00
900214	HM4	1.15	930616	HM4	0.75	961016	HM4	0.90
900314	HM4	1.10	930714	HM4	0.55	961113	HM4	0.50
900418	HM4	0.30	930818	HM4	0.95	961218	HM4	1.60
900516	HM4	1.35	930915	HM4	1.45	970115	HM4	0.95
900620	HM4	0.80	931013	HM4	1.80	970212	HM4	0.65
900718	HM4	0.95	931117	HM4	0.60	970312	HM4	1.40
900815	HM4	1.10	931215	HM4	0.95	970416	HM4	1.65
900912	HM4	1.00	940110	HM4	1.10	970514	HM4	1.25
901017	HM4	1.25	940214	HM4	0.40	970618	HM4	0.85
901114	HM4	1.50	940314	HM4	2.05	970716	HM4	1.20
901212	HM4	1.50	940411	HM4	1.25	970813	HM4	0.50
910116	HM4	1.60	940518	HM4	1.15	970917	HM4	0.80
910213	HM4	1.20	940613	HM4	1.15	971015	HM4	0.85
910320	HM4	1.50	940713	HM4	0.75	971112	HM4	1.45
910417	HM4	1.75	940817	HM4	0.75	971217	HM4	1.45
910515	HM4	1.10	940913	HM4	0.80	980114	HM4	0.75
910613	HM4	0.85	941011	HM4	0.80	980218	HM4	1.65
910717	HM4	0.95	941114	HM4	1.00	980318	HM4	1.25
910814	HM4	0.90	941212	HM4	1.20	980415	HM4	0.70
910918	HM4	1.25	950118	HM4	1.75	980513	HM4	0.50
911016	HM4	1.05	950213	HM4	1.80	980617	HM4	0.75
911113	HM4	1.60	950315	HM4	1.30	980722	HM4	1.00
911218	HM4	1.95	950412	HM4	1.40	980812	HM4	1.55
920115	HM4	1.90	950517	HM4	0.60	980916	HM4	0.75
920213	HM4	1.85	950614	HM4	1.15	981014	HM4	1.20
920319	HM4	2.25	950712	HM4	1.10	981125	HM4	0.55
920415	HM4	1.60	950816	HM4	0.55	981216	HM4	1.15

Appendix D2. New Zealand Data - Site HM6_NO3

Date	Site	NO3 (ppb)	Date	Site	NO3 (ppb)	Date	Site	NO3 (ppb)
890125	HM6	780	920513	HM6	710	950913	HM6	524
890222	HM6	445	920617	HM6	875	951011	HM6	504
890323	HM6	200	920715	HM6	765	951108	HM6	341
890419	HM6	220	920812	HM6	750	951213	HM6	244
890524	HM6	380	920917	HM6	585	960117	HM6	297
890614	HM6	690	921013	HM6	570	960214	HM6	117
890712	HM6	510	921120	HM6	440	960313	HM6	92
890817	HM6	810	921218	HM6	605	960417	HM6	430
890913	HM6	1135	930113	HM6	150	960515	HM6	405
891018	HM6	695	930217	HM6	220	960612	HM6	507
891116	HM6	390	930317	HM6	160	960717	HM6	495
891214	HM6	370	930414	HM6	185	960814	HM6	820
900117	HM6	175	930512	HM6	550	960918	HM6	832
900214	HM6	555	930616	HM6	840	961016	HM6	471
900314	HM6	265	930714	HM6	685	961113	HM6	387
900418	HM6	285	930818	HM6	580	961218	HM6	355
900516	HM6	360	930915	HM6	425	970115	HM6	471
900620	HM6	485	931013	HM6	455	970212	HM6	333
900718	HM6	790	931117	HM6	340	970312	HM6	606
900815	HM6	855	931215	HM6	325	970416	HM6	377
900912	HM6	670	940112	HM6	218	970514	HM6	380
901017	HM6	575	940216	HM6	3	970618	HM6	542
901114	HM6	475	940316	HM6	202	970716	HM6	777
901212	HM6	215	940413	HM6	1188	970813	HM6	440
910116	HM6	8	940518	HM6	409	970917	HM6	704
910213	HM6	9	940615	HM6	696	971015	HM6	335
910320	HM6	160	940713	HM6	742	971112	HM6	380
910417	HM6	185	940817	HM6	723	971217	HM6	214
910515	HM6	500	940914	HM6	359	980114	HM6	41
910613	HM6	210	941012	HM6	749	980218	HM6	87
910717	HM6	520	941116	HM6	442	980318	HM6	373
910814	HM6	820	941214	HM6	256	980415	HM6	481
910918	HM6	605	950118	HM6	40	980513	HM6	451
911016	HM6	480	950215	HM6	51	980617	HM6	1091
911113	HM6	350	950315	HM6	570	980722	HM6	987
911218	HM6	92	950412	HM6	1052	980812	HM6	892
920115	HM6	105	950517	HM6	438	980916	HM6	387
920213	HM6	415	950614	HM6	944	981014	HM6	535
920319	HM6	110	950712	HM6	928	981125	HM6	544
920415	HM6	395	950816	HM6	792	981216	HM6	371

Appendix D3. New Zealand Data - Site RO1_NH4

Date	Site	NH4 (ppb)	Date	Site	NH4 (ppb)	Date	Site	NH4 (ppb)
890215	RO1	11	920617	RO1	2	951011	RO1	2
890315	RO1	5	920715	RO1	7	951115	RO1	3
890412	RO1	4	920812	RO1	3	951212	RO1	2
890510	RO1	2	920916	RO1	3	960117	RO1	3
890615	RO1	3	921014	RO1	5	960215	RO1	5
890719	RO1	6	921112	RO1	2	960312	RO1	4
890815	RO1	3	921209	RO1	5	960417	RO1	5
890913	RO1	5	930113	RO1	8	960515	RO1	1
891011	RO1	6	930217	RO1	6	960612	RO1	1
891115	RO1	9	930316	RO1	2	960718	RO1	5
891212	RO1	5	930414	RO1	2	960814	RO1	2
900117	RO1	4	930512	RO1	3	960912	RO1	2
900214	RO1	3	930615	RO1	6	961016	RO1	4
900314	RO1	2	930715	RO1	4	961112	RO1	3
900418	RO1	4	930811	RO1	3	961212	RO1	4
900516	RO1	5	930915	RO1	4	970116	RO1	1
900613	RO1	4	931014	RO1	2	970212	RO1	4
900718	RO1	3	931117	RO1	3	970311	RO1	5
900815	RO1	5	931215	RO1	4	970416	RO1	1
900912	RO1	2	940113	RO1	.	970515	RO1	2
901017	RO1	1	940216	RO1	.	970611	RO1	1
901114	RO1	6	940316	RO1	.	970716	RO1	2
901212	RO1	2	940413	RO1	.	970812	RO1	0
910116	RO1	3	940511	RO1	.	970917	RO1	1
910213	RO1	10	940615	RO1	.	971015	RO1	1
910312	RO1	11	940713	RO1	.	971113	RO1	1
910416	RO1	8	940817	RO1	.	971218	RO1	1
910515	RO1	1	940914	RO1	.	980114	RO1	2
910612	RO1	2	941012	RO1	.	980211	RO1	2
910717	RO1	1	941117	RO1	.	980311	RO1	3
910814	RO1	2	941214	RO1	.	980415	RO1	0
910911	RO1	9	950111	RO1	3	980513	RO1	2
911016	RO1	1	950215	RO1	3	980618	RO1	4
911113	RO1	2	950315	RO1	2	.	RO1	.
911212	RO1	4	950411	RO1	3	980812	RO1	4
920115	RO1	2	950517	RO1	1	980916	RO1	3
920212	RO1	2	950614	RO1	2	981014	RO1	2
920318	RO1	1	950713	RO1	4	981111	RO1	1
920415	RO1	1	950816	RO1	2	981217	RO1	2
920513	RO1	5	950914	RO1	2			

Appendix D4. New Zealand Data - Site RO2_NH4

Date	Site	NH4 (ppb)	Date	Site	NH4 (ppb)	Date	Site	NH4 (ppb)
890215	RO2	48	920617	RO2	74	951011	RO2	64
890315	RO2	38	920715	RO2	57	951115	RO2	48
890412	RO2	46	920812	RO2	75	951212	RO2	66
890510	RO2	55	920916	RO2	69	960117	RO2	58
890615	RO2	54	921014	RO2	7	960215	RO2	69
890719	RO2	37	921112	RO2	75	960312	RO2	43
890815	RO2	7	921209	RO2	62	960417	RO2	64
890913	RO2	57	930113	RO2	65	960515	RO2	47
891011	RO2	16	930217	RO2	62	960612	RO2	15
891115	RO2	35	930316	RO2	9	960718	RO2	35
891212	RO2	46	930414	RO2	63	960814	RO2	20
900117	RO2	33	930512	RO2	69	960912	RO2	22
900214	RO2	36	930615	RO2	62	961016	RO2	59
900314	RO2	31	930715	RO2	49	961112	RO2	48
900418	RO2	52	930811	RO2	75	961212	RO2	54
900516	RO2	31	930915	RO2	57	970116	RO2	70
900613	RO2	43	931014	RO2	81	970212	RO2	65
900718	RO2	8	931117	RO2	67	970311	RO2	62
900815	RO2	33	931215	RO2	61	970416	RO2	60
900912	RO2	27	940113	RO2	.	970515	RO2	53
901017	RO2	23	940216	RO2	.	970611	RO2	68
901114	RO2	38	940316	RO2	.	970716	RO2	41
901212	RO2	22	940413	RO2	.	970812	RO2	47
910116	RO2	49	940511	RO2	.	970917	RO2	59
910213	RO2	64	940615	RO2	.	971015	RO2	52
910312	RO2	38	940713	RO2	.	971113	RO2	54
910416	RO2	43	940817	RO2	.	971218	RO2	34
910515	RO2	42	940914	RO2	.	980115	RO2	47
910612	RO2	44	941012	RO2	.	980211	RO2	47
910717	RO2	58	941117	RO2	.	980311	RO2	66
910814	RO2	52	941214	RO2	.	980415	RO2	33
910911	RO2	66	950111	RO2	77	980513	RO2	52
911016	RO2	34	950215	RO2	71	980618	RO2	45
911113	RO2	57	950315	RO2	55	980714	RO2	58
911212	RO2	78	950411	RO2	45	980812	RO2	50
920115	RO2	51	950517	RO2	60	980916	RO2	53
920212	RO2	81	950614	RO2	97	981014	RO2	54
920318	RO2	42	950713	RO2	71	981111	RO2	40
920415	RO2	56	950816	RO2	81	981217	RO2	28
920513	RO2	53	950914	RO2	46			

Appendix E.
Algorithm for Interval Testing in MS-Excel

Appendix E: Algorithm for Interval Testing in MS-Excel (McBride, 1999b)

Example calculations, Waimea Creek data, using Excel™

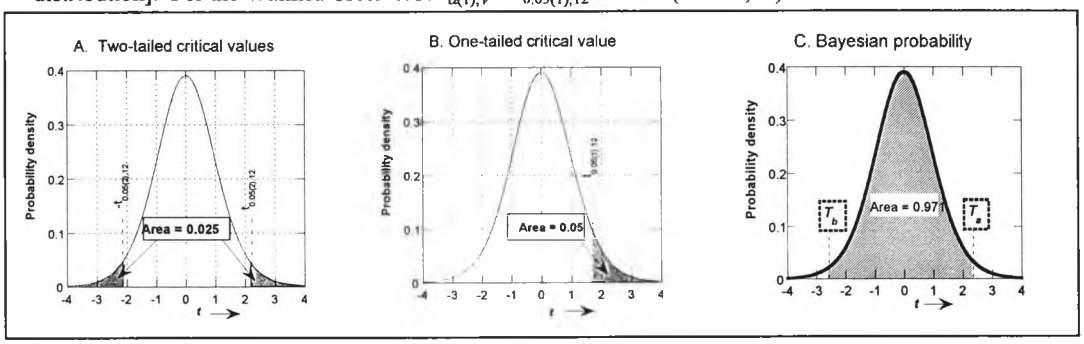
Input data

- Upstream taxonomic richness (per 0.1 m^2 sampling area, 7 replicates): 13, 12, 18, 11, 14, 11, 13
- Downstream taxonomic richness (7 replicates): 12, 14, 12, 11, 15, 13, 14
- Significance level for each comparison: $\alpha = 5\%$.
- Lower and upper bounds on environmentally significant percentage change from the upstream mean taxonomic richness: $D_L = -20\%$, $D_U = +20\%$.

Calculated degrees of freedom and critical *t* values for hypothesis tests

With $n_{up} = n_{down} = 7$ replicates at each site there are $\nu = n_{up} + n_{down} - 2 = 12$ degrees of freedom for each comparison. Critical values ("inverses") of the *t*-distribution are calculated using Excel's TINV function, which gives the value of *t* that cuts off a given total area in *both* tails of the distribution.

- For the null hypothesis test we need $t_{\alpha(2),\nu}$ ["$\alpha(2)$" denotes the two-tailed value, cutting off an area $\frac{1}{2}\alpha$ in each tail of the distribution]. For the Waimea Creek case $t_{\alpha(2),\nu} = t_{0.05(2),12}$ = TINV(0.05,12) = 2.179.
- For equivalence tests we need $t_{\alpha(1),\nu}$ ["$\alpha(1)$" signifies a one-tailed value, cutting off an area α in the upper tail of the distribution]. For the Waimea Creek case $t_{\alpha(1),\nu} = t_{0.05(1),12}$ = TINV(2*0.05,12) = 1.782.



Derived data

Required for all procedures

- upstream & downstream means; estimated difference: $\bar{x}_{up} = 13.14$, $\bar{x}_{down} = 13.00$; $\hat{d} = \bar{x}_{down} - \bar{x}_{up} = -0.14$
- upstream and downstream standard deviations: $s_{up} = 2.41$, $s_{down} = 1.41$
- pooled standard deviation: $s_p = \sqrt{\left(n_{up}s_{up}^2 + n_{down}s_{down}^2\right)/\left(n_{up} + n_{down}\right)} = 1.98$
- standard error: $SE = s_p\sqrt{1/n_{up} + 1/n_{down}} = 1.06$

Required for null hypothesis test

- test statistic: $T = |\hat{d}|/SE = 0.14$

Required for equivalence tests

- lower equivalence interval limit: $d_L = \bar{x}_{up}D_L/100 = -2.63$
- upper equivalence interval limit: $d_U = \bar{x}_{up}D_U/100 = 2.63$
- lower test statistic: $T_a = \left(\hat{d} - d_L\right)/SE = 2.35$
- upper test statistic: $T_b = \left(\hat{d} - d_U\right)/SE = -2.62$

Required for Bayesian calculations

The preceding four items, plus:

- cumulative *t* probability up to T_a: $F(T_a) = 98.2\%$
- cumulative *t* probability up to T_b: $F(T_b) = 1.1\%$

where $F(t)$ is the cumulative *t*-distribution probability, calculated using Excel's TDIST function via the formula $F(t) = \frac{1}{2} + \text{SIGN}(t)(\frac{1}{2} - \text{TDIST}(\text{ABS}(t),\nu,1))$. (The formula accounts for cases where *t* is negative.)

Appendix E. Algorithm for Interval Testing in MS-Excel (McBride, 1999b)

Rules and outcomes

- The null hypothesis is rejected if $T > t_{\alpha(2),\nu}$. This condition is not satisfied (because 0.14 < 2.179) and so we do not reject the hypothesis. Therefore we infer "no statistically significant difference".
- The equivalence hypothesis is rejected if *either* $T_a < -t_{\alpha(1),\nu}$ or $T_b > t_{\alpha(1),\nu}$. Neither condition is satisfied (because 2.35 > –1.782 and –2.62 < 1.782) and so the hypothesis is not rejected. Therefore we infer "equivalence".
- The inequivalence hypothesis is rejected if *both* $T_a \geq t_{\alpha(1),\nu}$ and $T_b \leq t_{\alpha(1),\nu}$. Both conditions are satisfied (because 2.35 > 1.782 and –2.62 < 1.782) and so the hypothesis is rejected. Therefore we infer "equivalence".
- The Bayesian probability that the true difference lies within the equivalence interval is $F(T_a) - F(T_b) = 97.1\%$

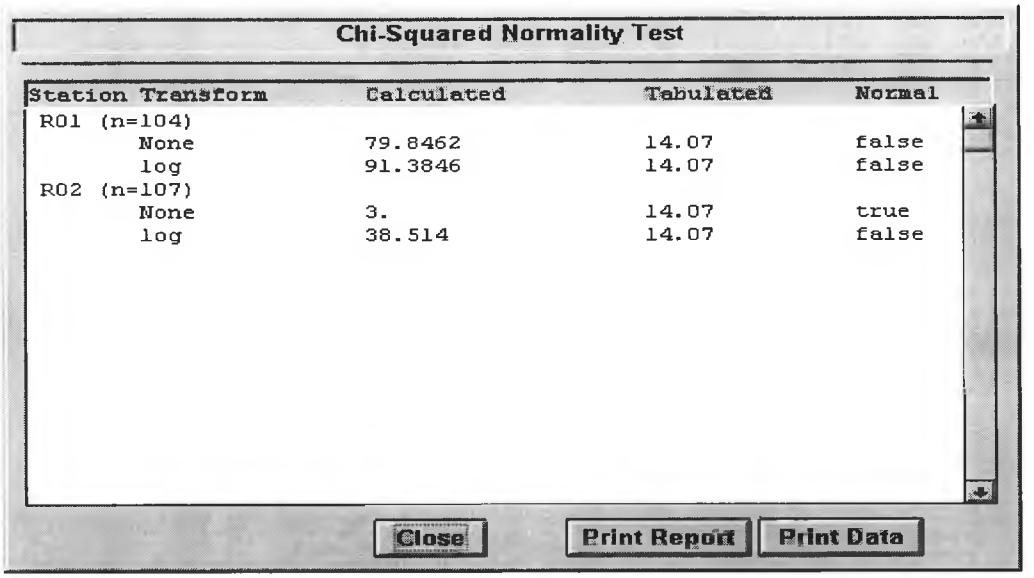
Multiple comparisons

The null and equivalence hypothesis tests have used a significance level of $\alpha = 5\%$. This means that the risk of falsely rejecting a true hypothesis is 5% *for each comparison* (i.e., for each stream). To keep the risk over *all* comparisons to 5% one must adjust the significance level downward. The pessimistic (Bonferroni) adjustment reduces α to 0.85%. In that case $t_{\alpha(1),\nu} = 2.769$ so that the inequivalence hypothesis would *not* be rejected.

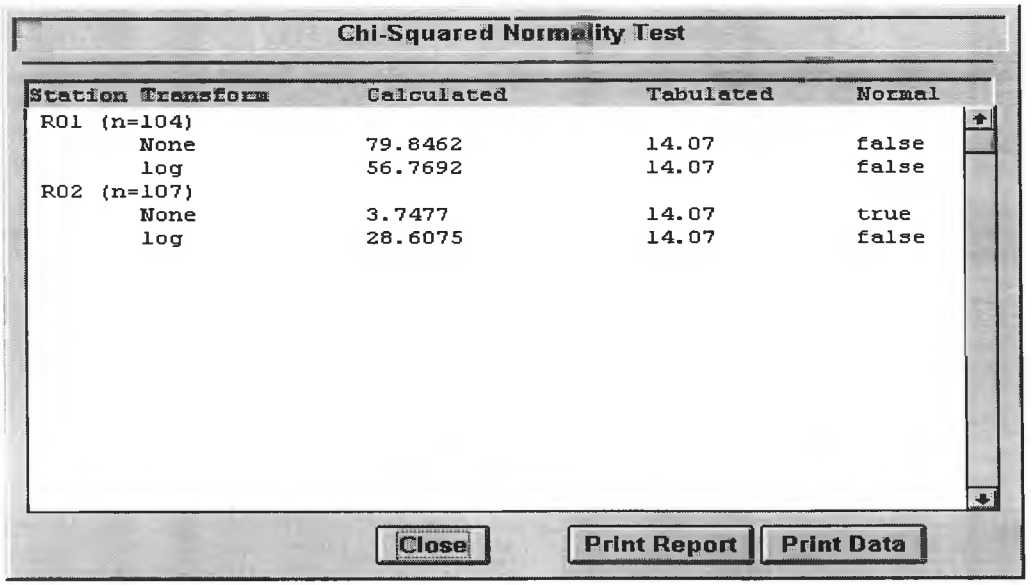
Appendix F.
Normality Test Results in WQStat Plus™

Appendix F. Normality Test Results in WQStat Plus™

Raw NH4 data for RO1 and RO2

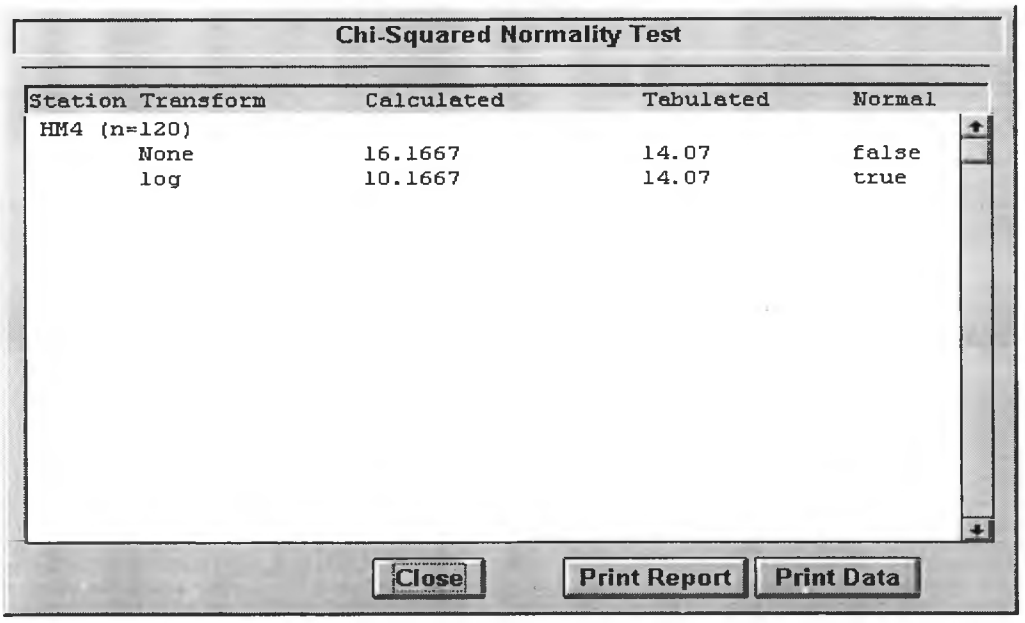


Flow-adjusted NH4 data for RO1 and RO2

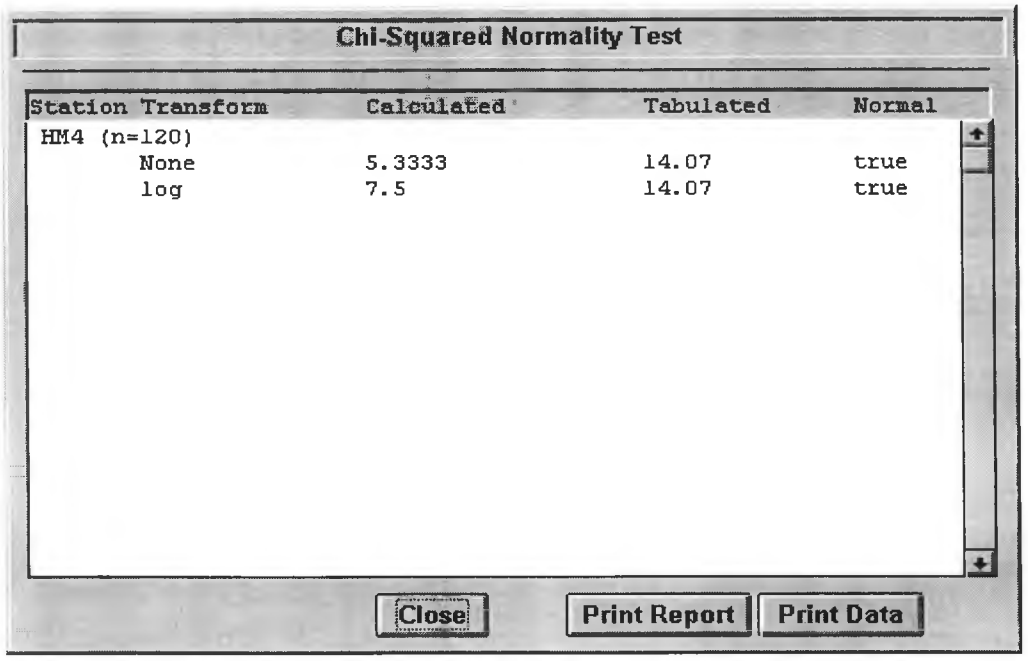


Appendix F. Normality Test Results in WQStat Plus™

Raw BOD5 data for HM4

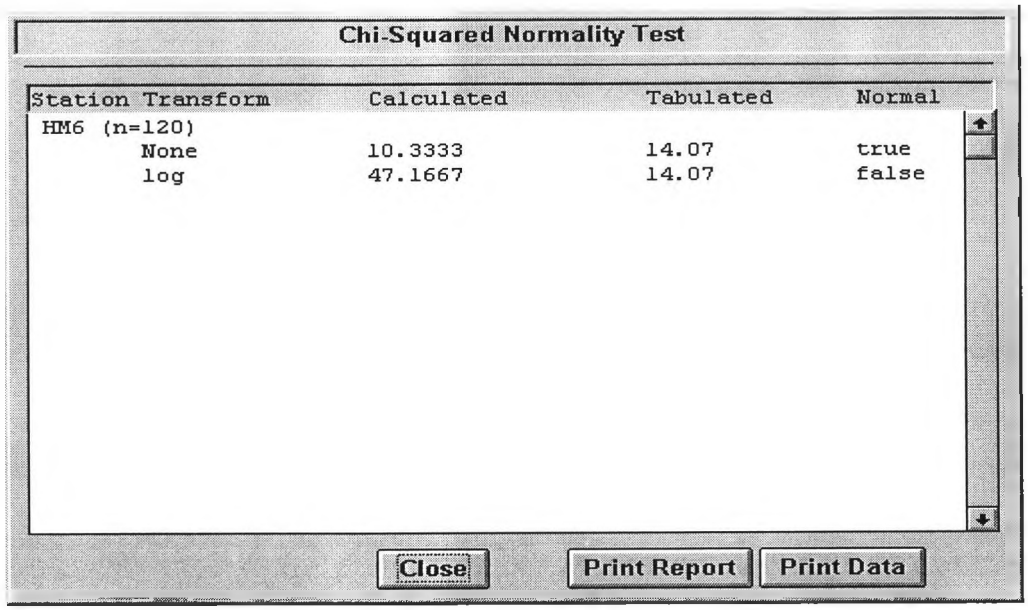


Flow-adjusted BOD5 data for HM4

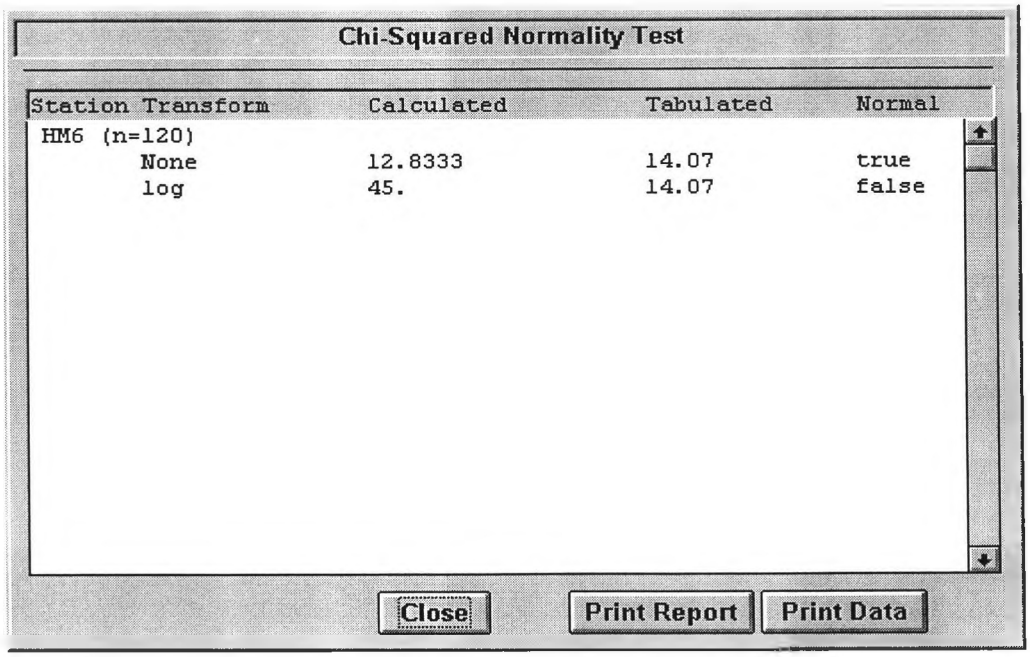


Appendix F. Normality Test Results in WQStat Plus™

Raw NO3 data for HM6



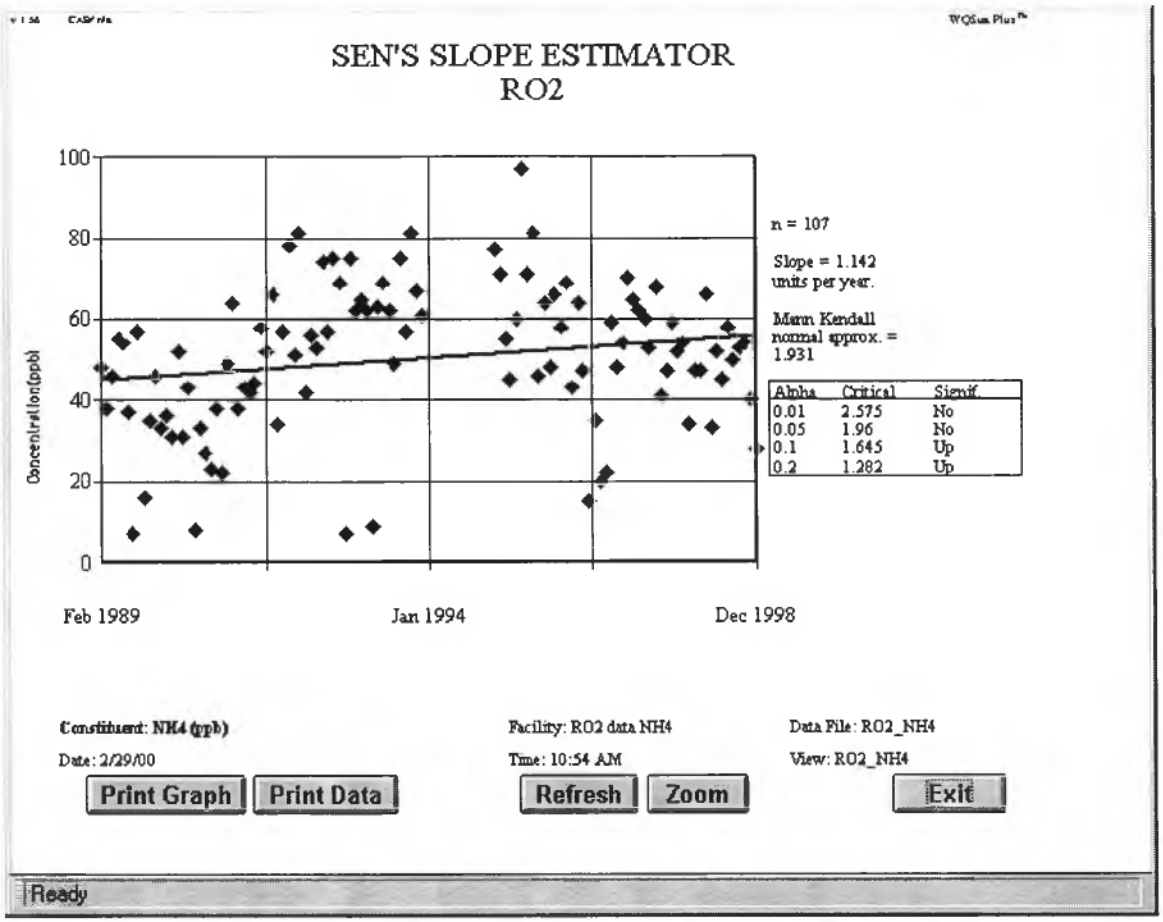
Flow-adjusted NO3 data for HM6



Appendix G.
Trend Analysis Results in WQStat Plus™

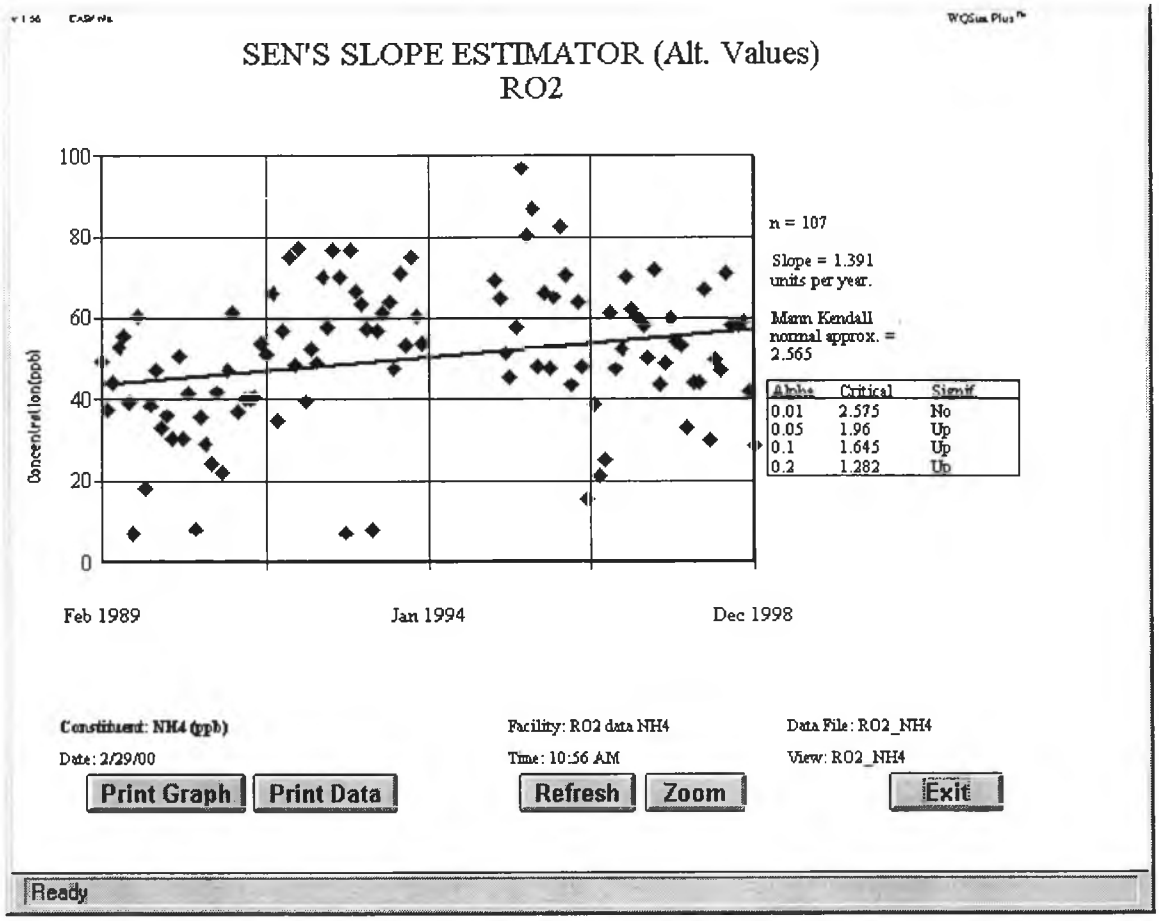
Appendix G1. Mann-Kendall Results for RO2_NH4

10-yr, raw data



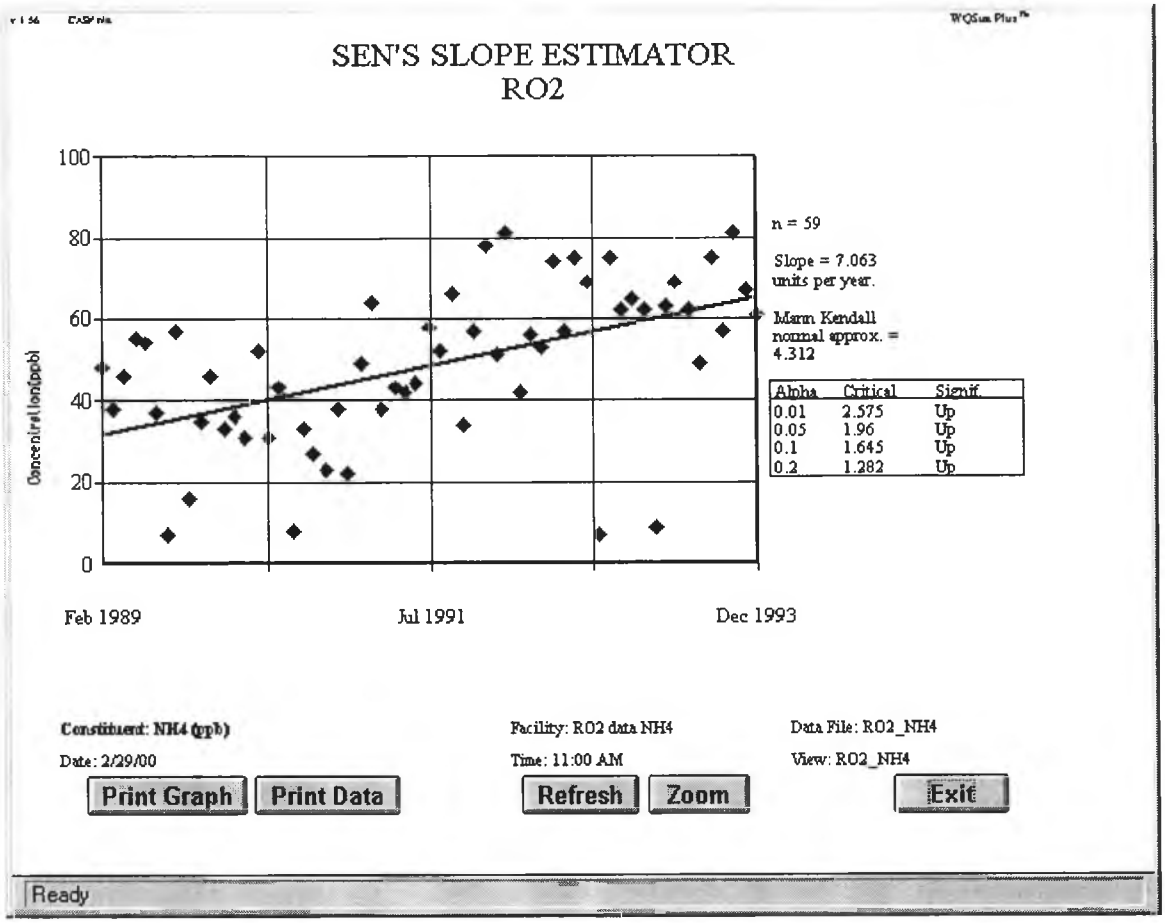
Appendix G1. Mann-Kendall Results for RO2_NH4

10 yr, FAC



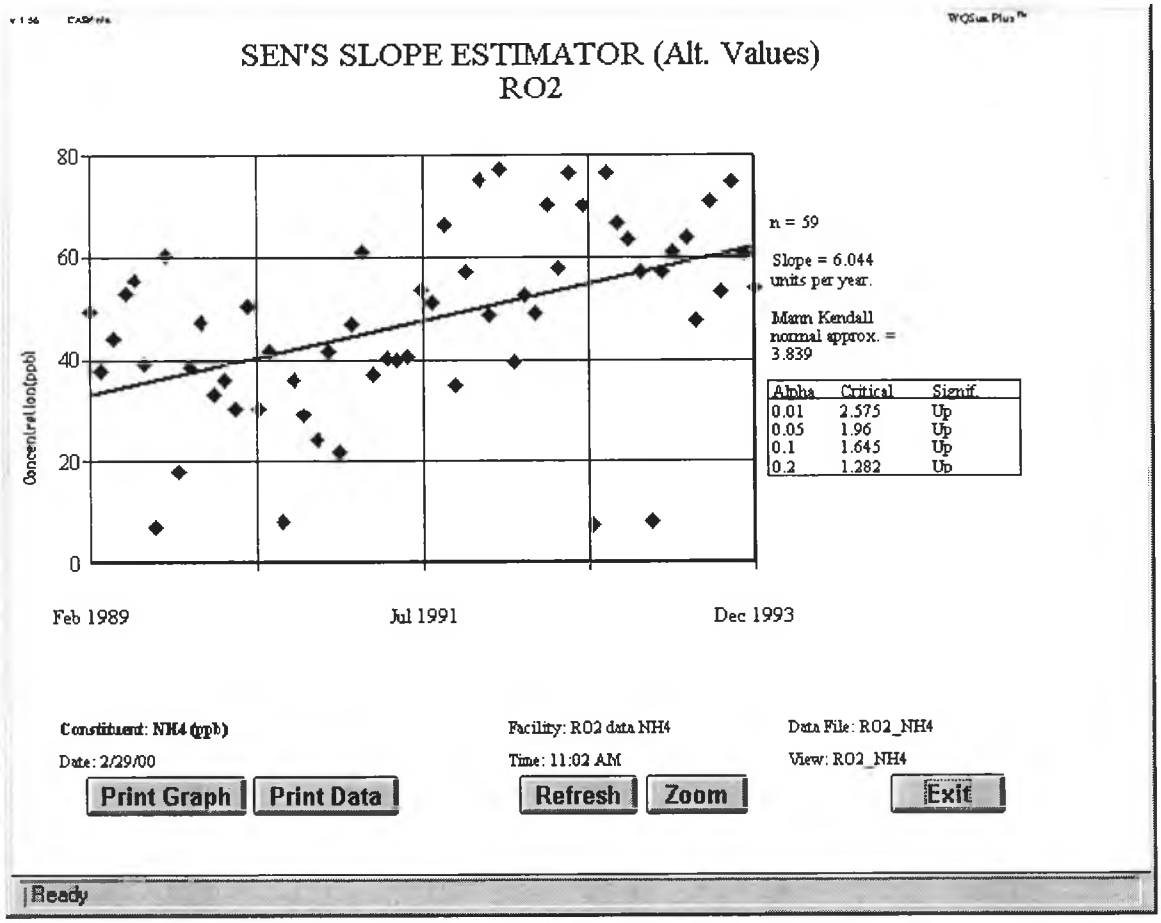
Appendix G1. Mann-Kendall Results for RO2-NH4

1st 5-yr, raw



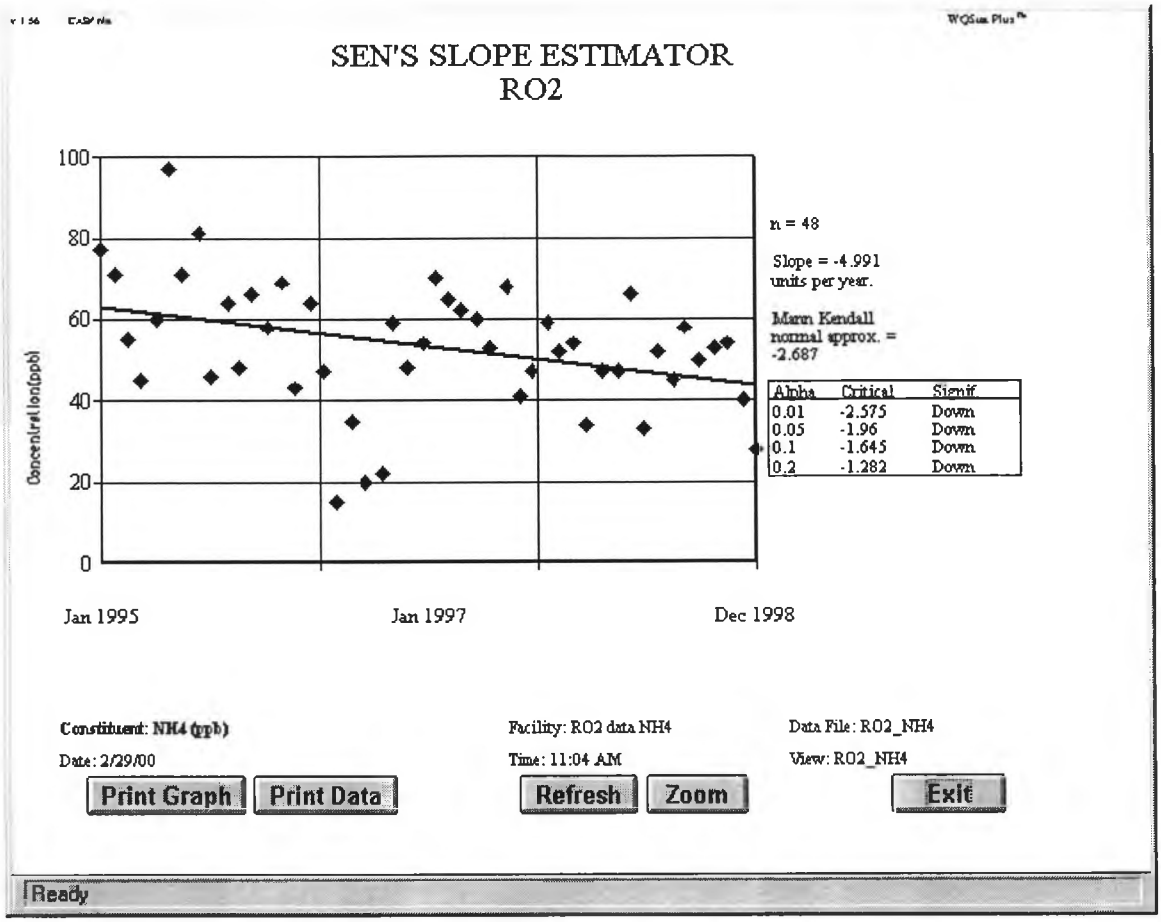
Appendix G1. Mann-Kendall Results for RO2_NH4

1st 5-yr, FAC



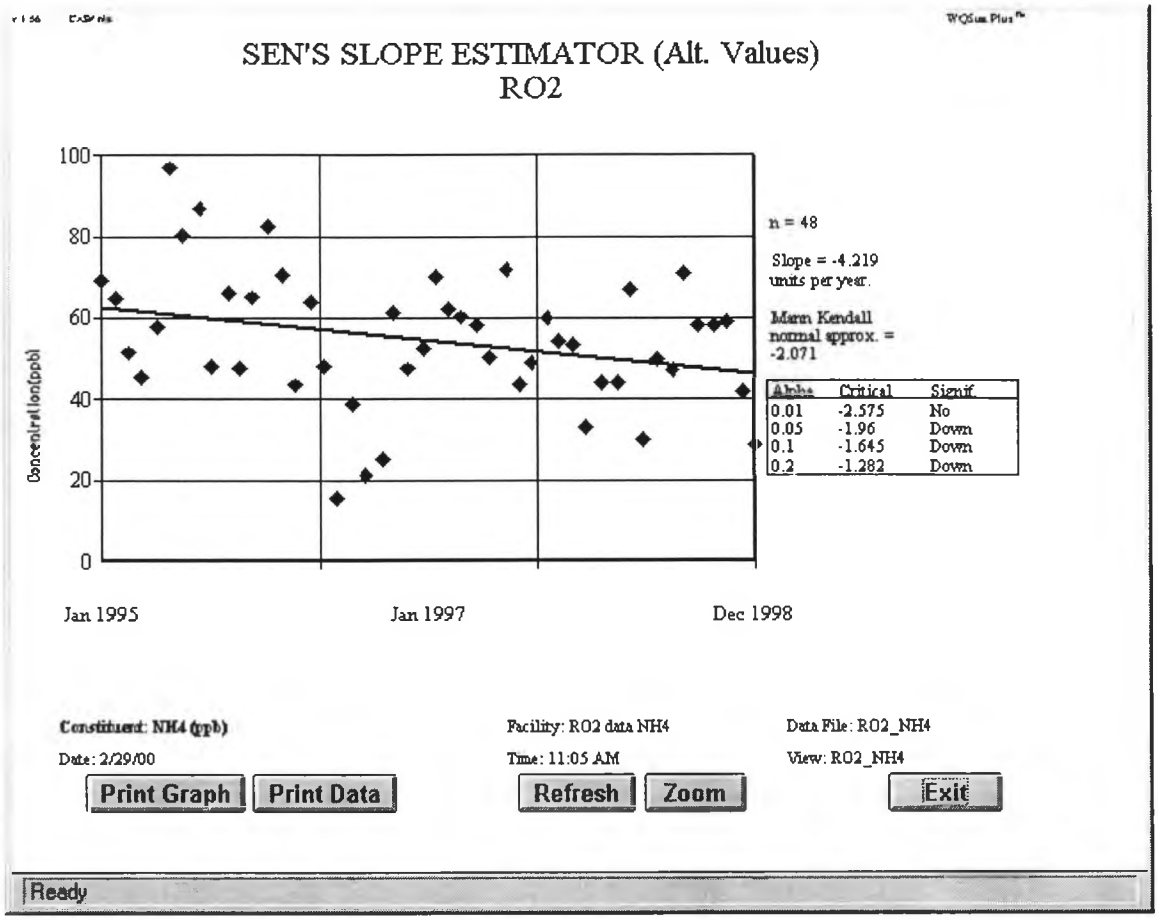
Appendix G1. Mann-Kendall Results for RO2_NH4

2nd 5-yr, raw



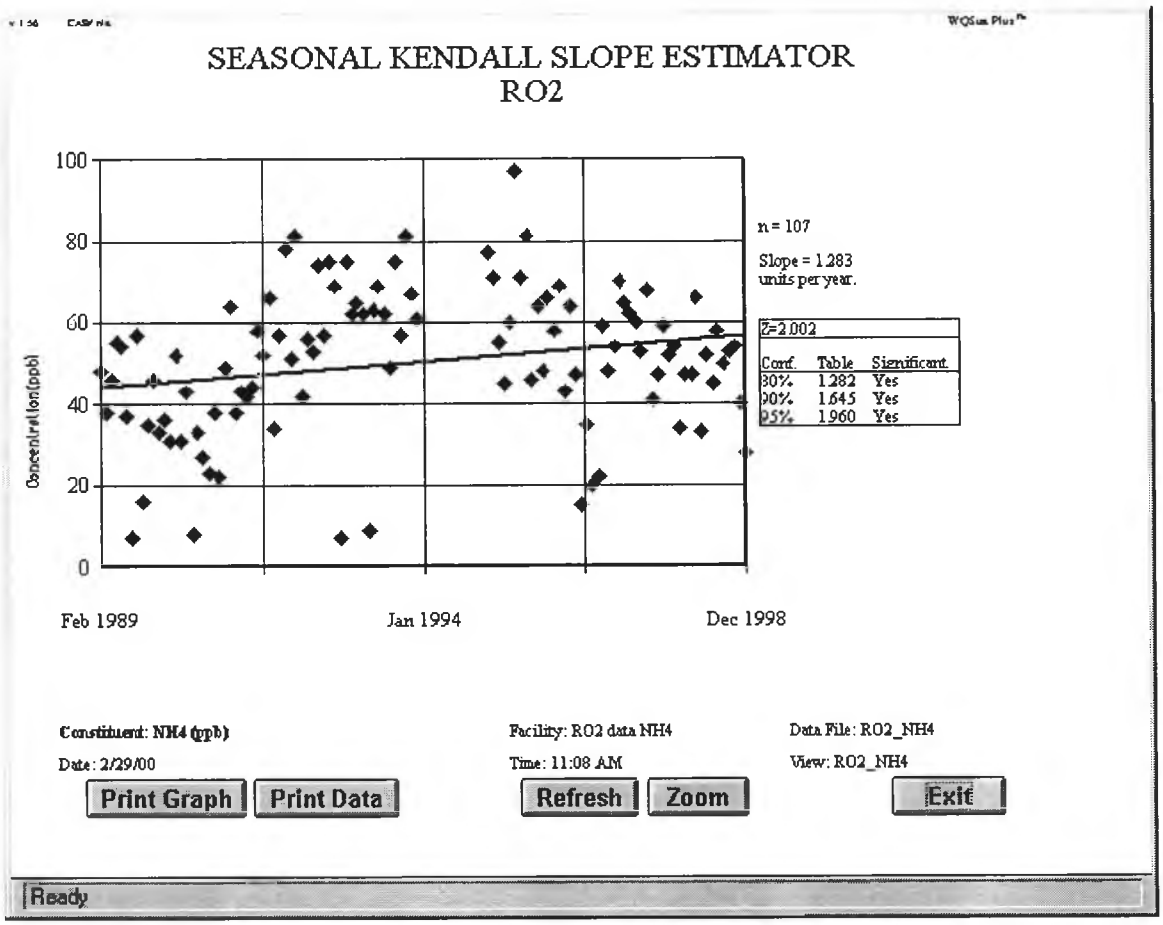
Appendix G1. Mann-Kendall Results for RO2_NH4

2nd 5-yr, FAC



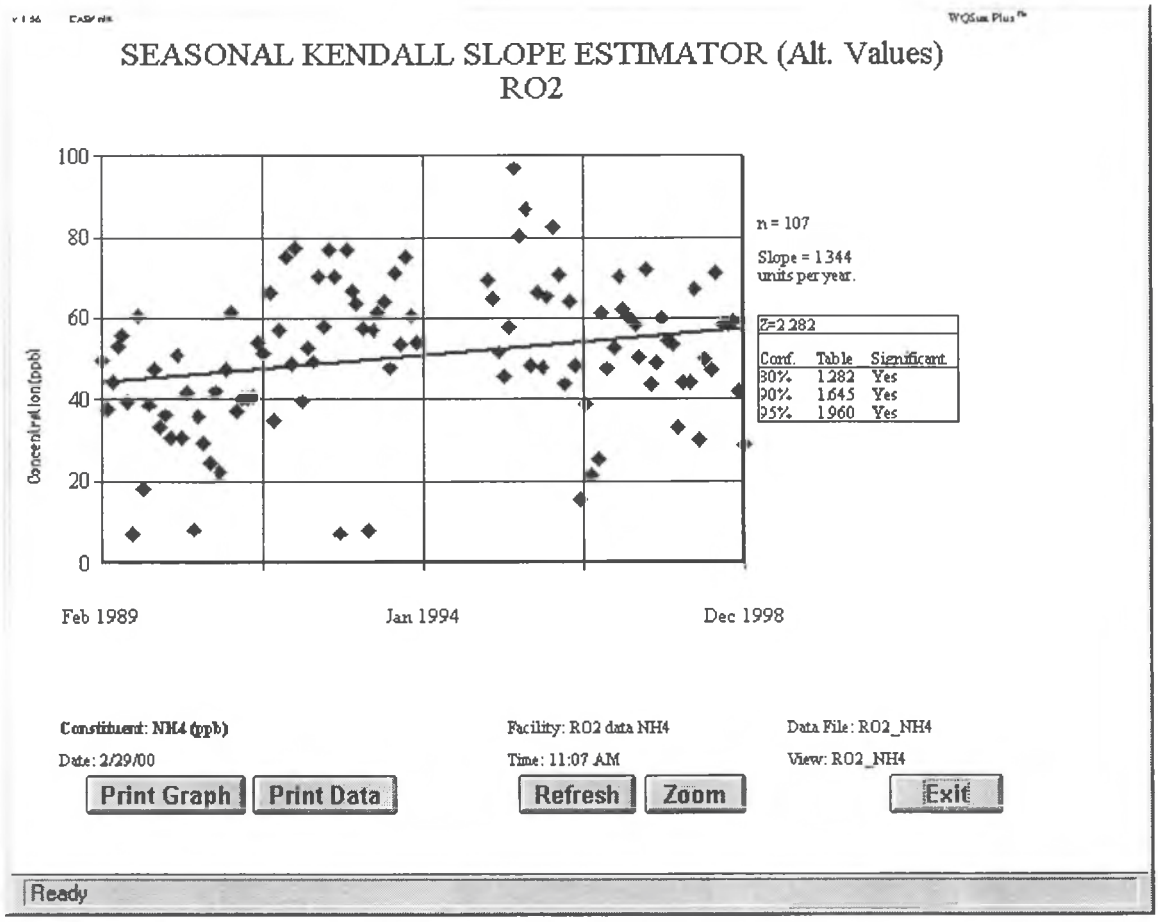
Appendix G2. Seasonal Kendall Results for RO2_NH4

10-yr, raw



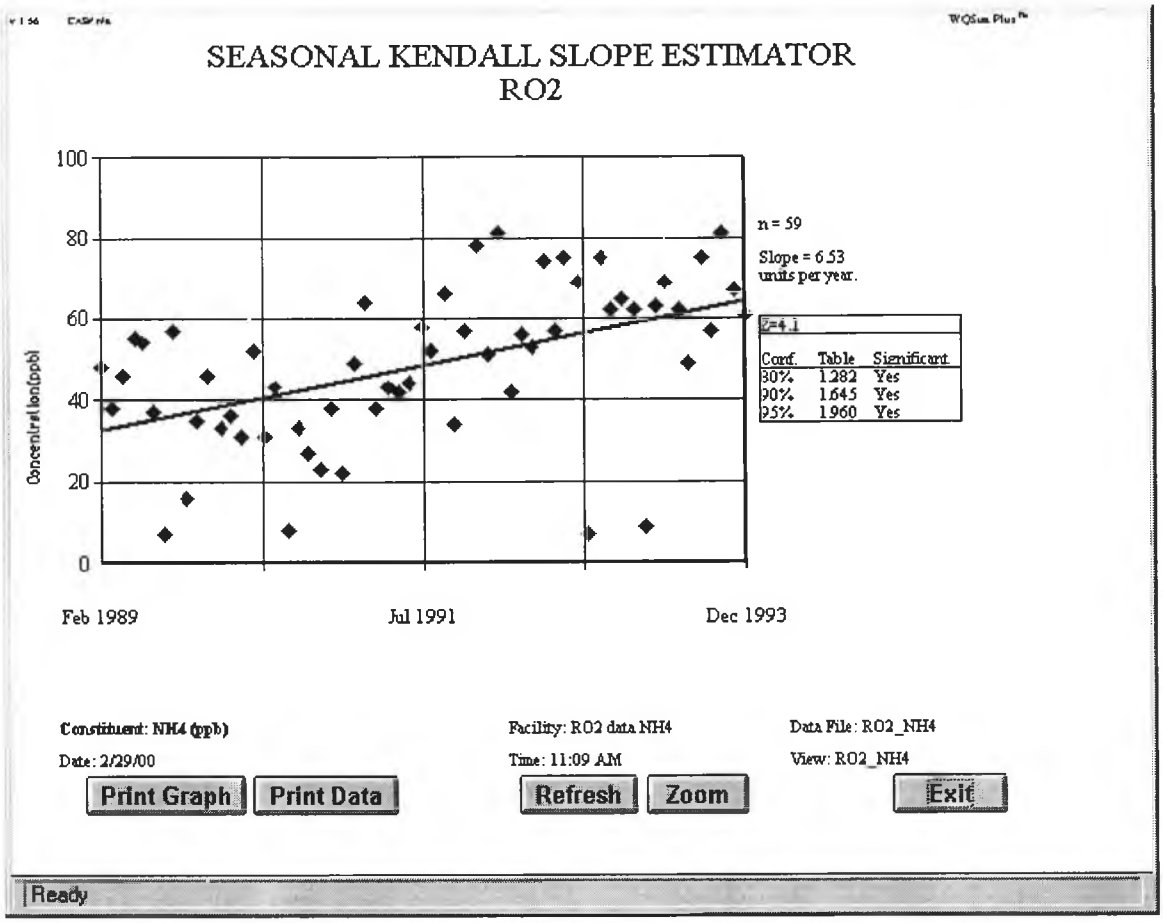
Appendix G2. Seasonal Kendall Results for RO2_NH4

10-yr, FAC



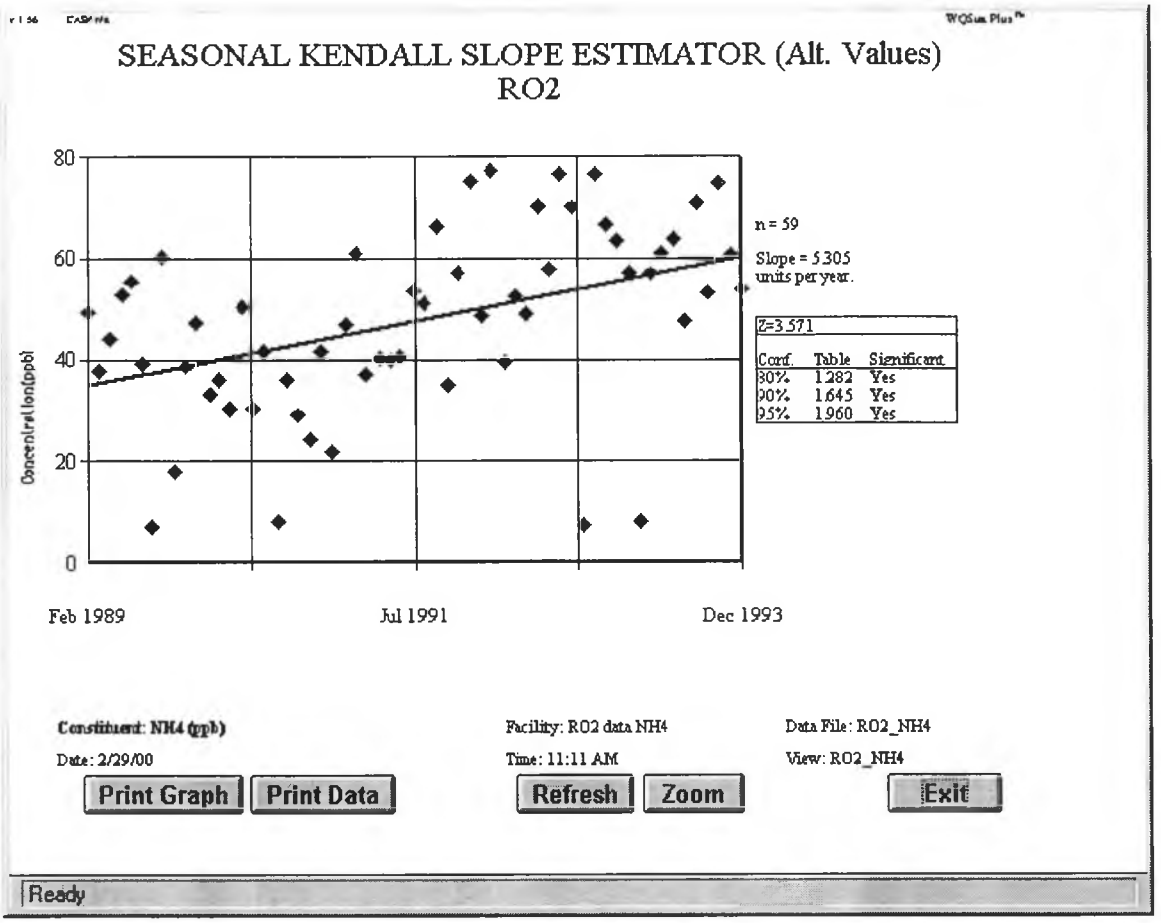
Appendix G2. Seasonal Kendall Results for RO2_NH4

1st 5-yr, raw



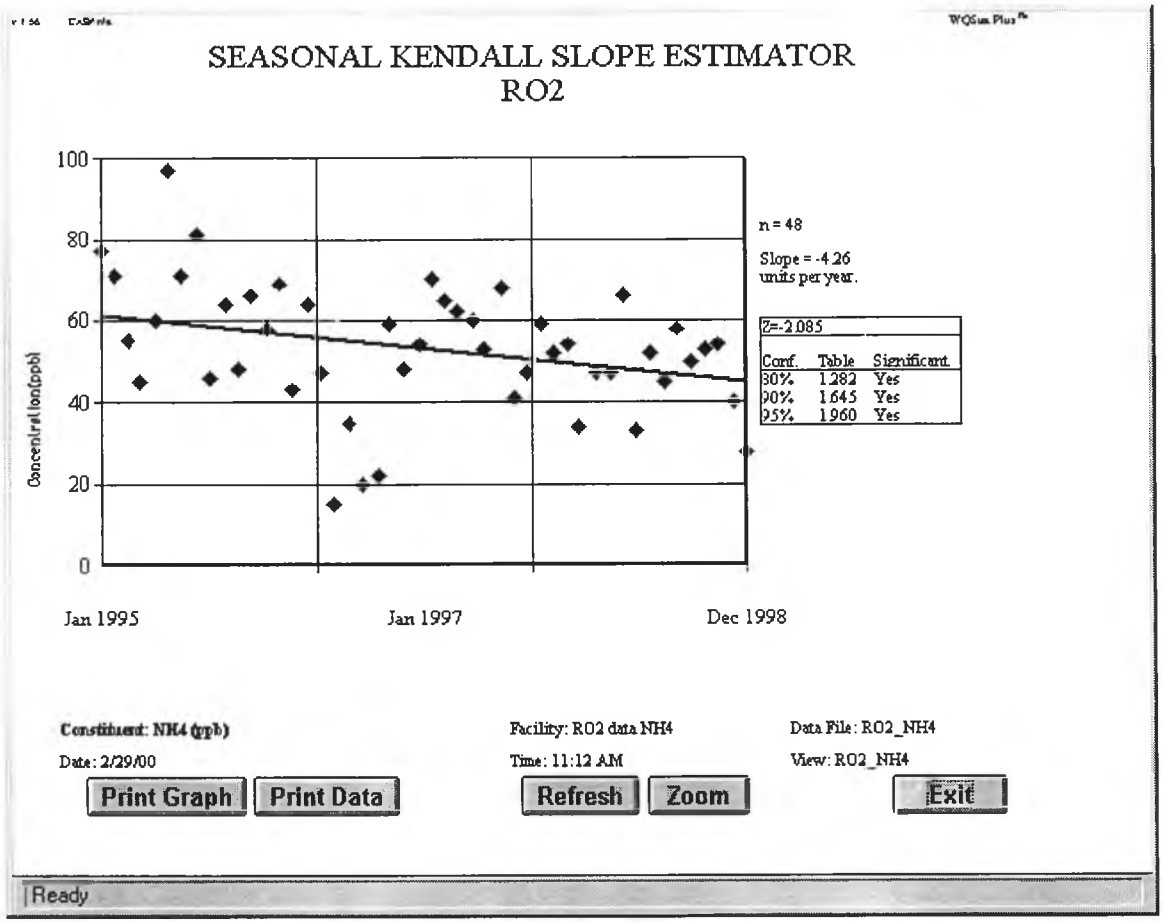
Appendix G2. Seasonal Kendall Results for RO2_NH4

1st 5-yr, FAC



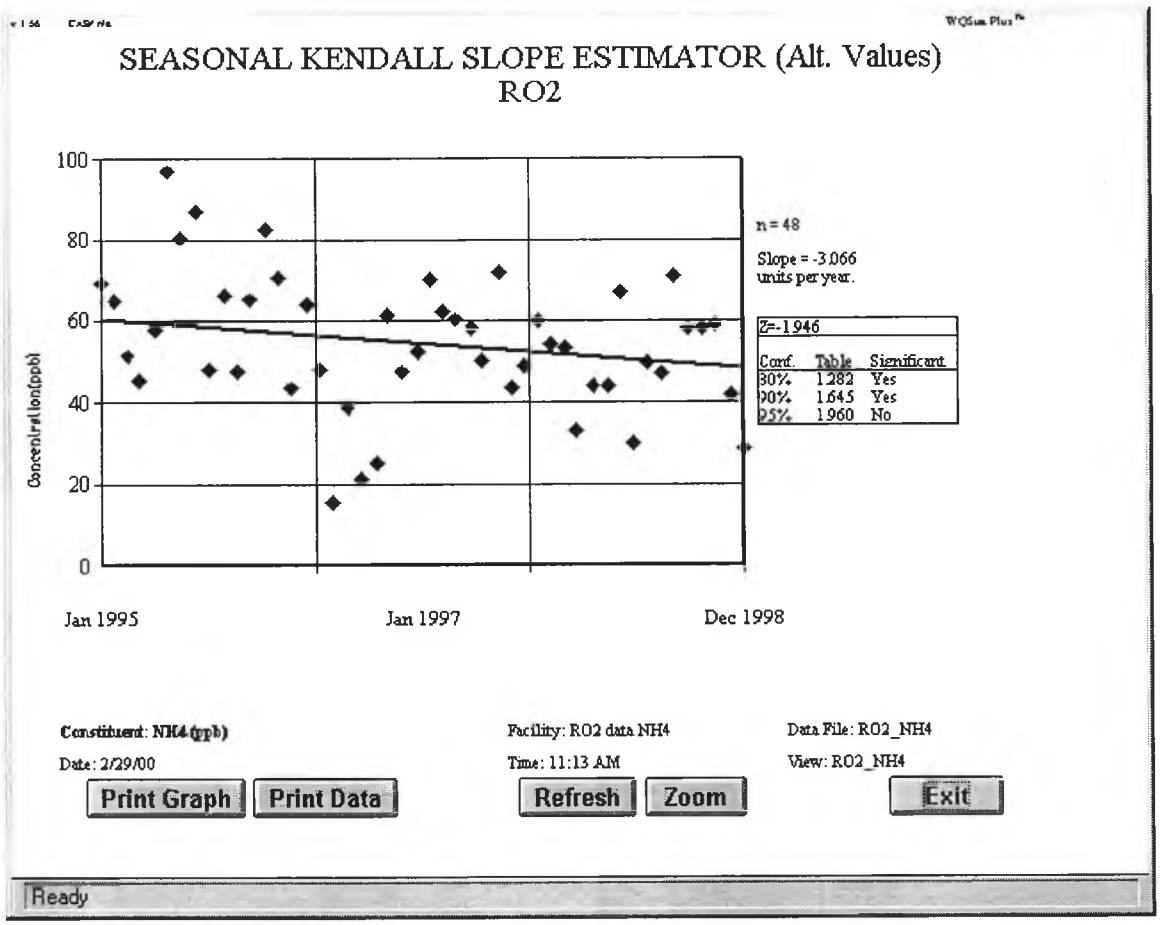
Appendix G2. Seasonal Kendall Results for RO2_NH4

2nd 5-yr, raw



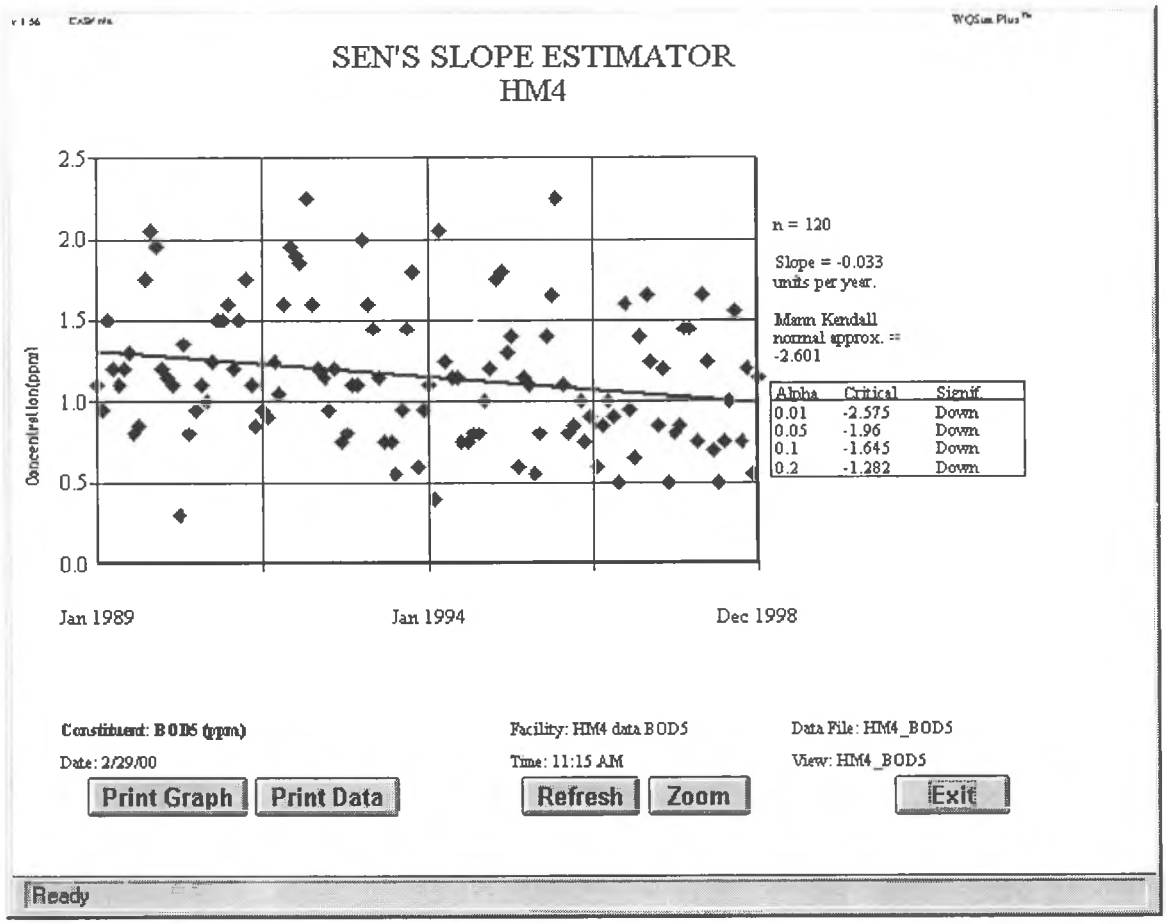
Appendix G2. Seasonal Kendall Results for RO2_NH4

2nd 5-yr, FAC



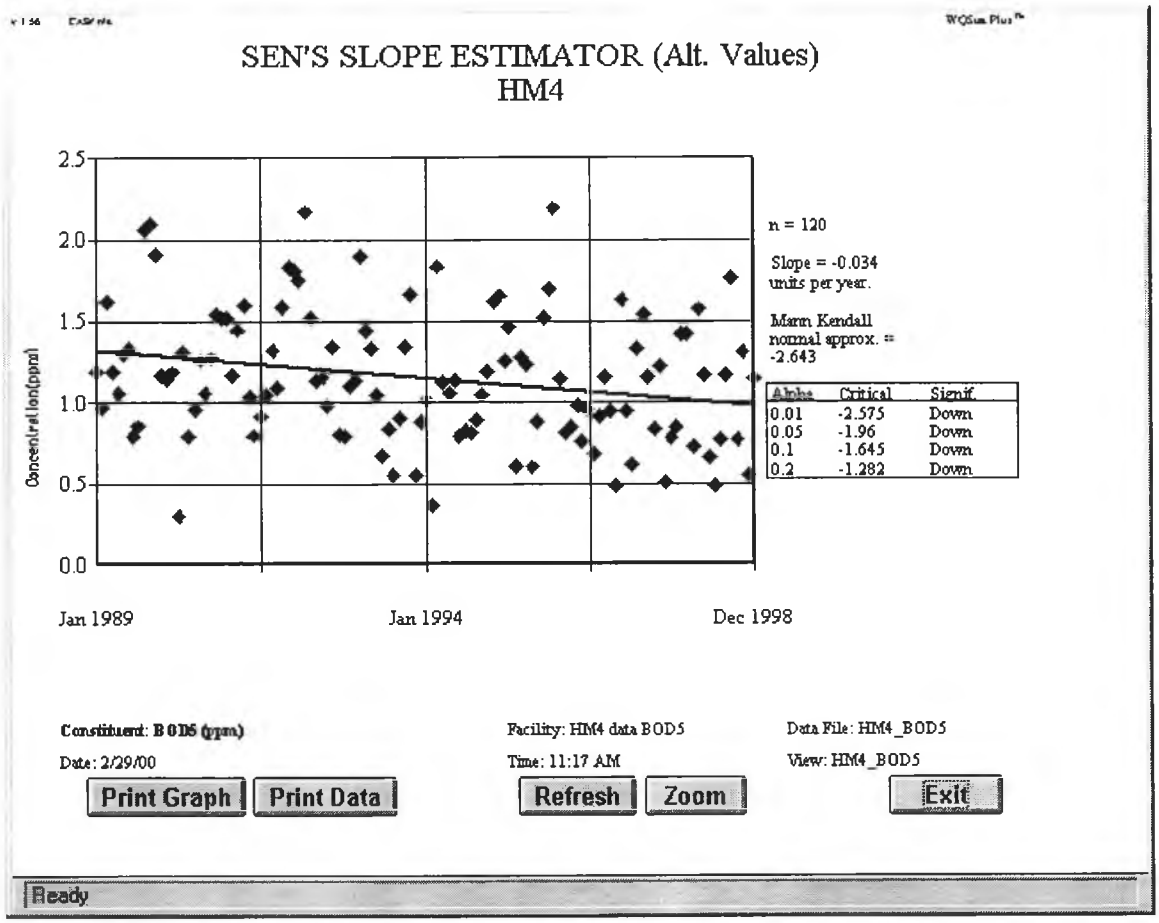
Appendix G3. Mann-Kendall Results for HM4_BOD5

10 yr, raw



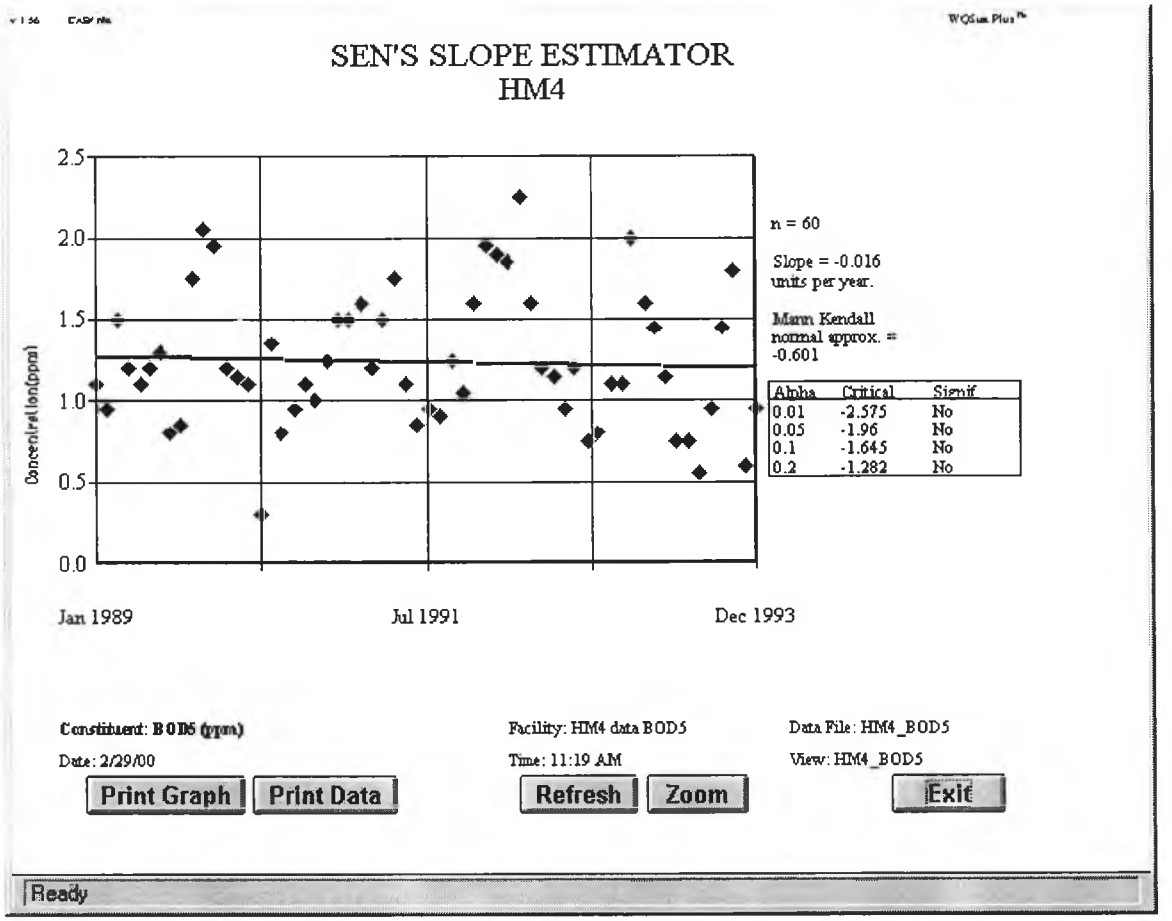
Appendix G3. Mann-Kendall Results for HM4_BOD5

10-yr, FAC



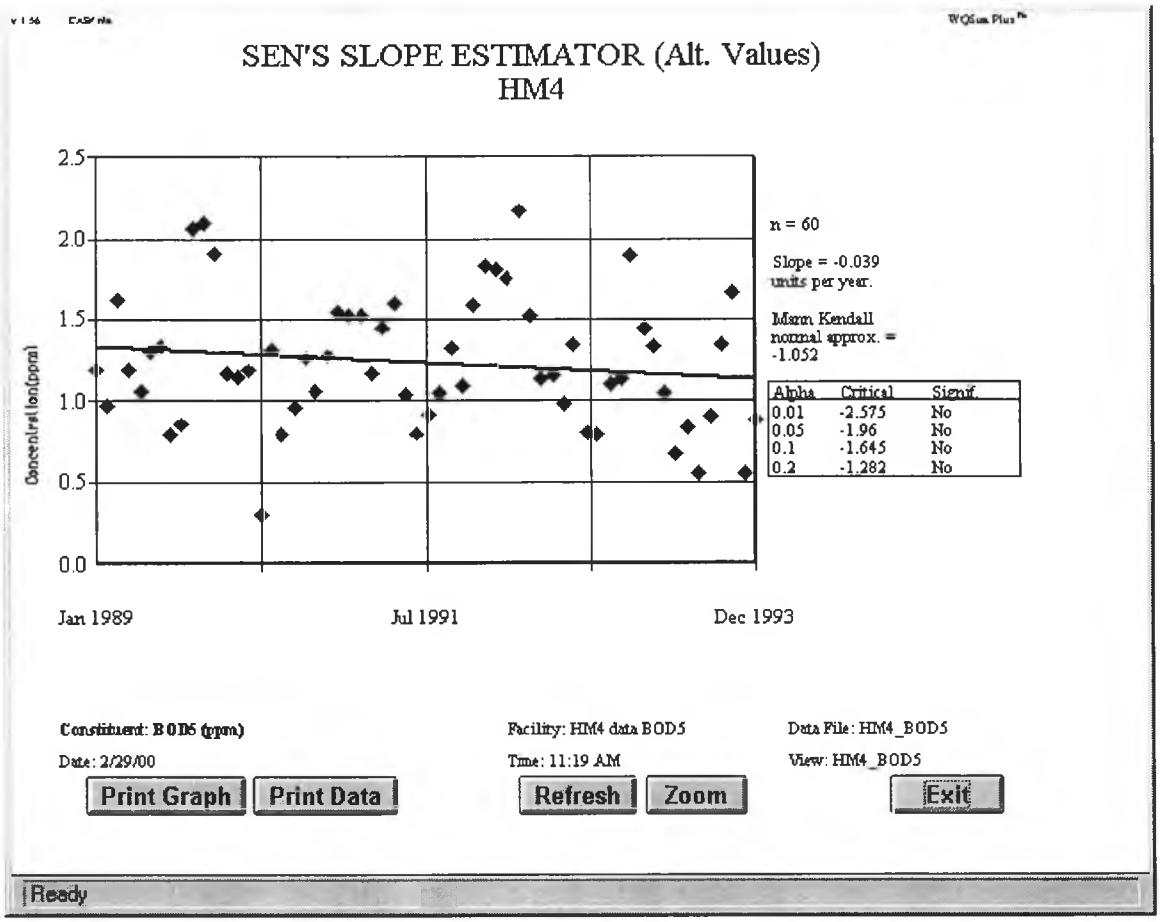
Appendix G3. Mann-Kendall Results for HM4_BOD5

1st 5-yr, raw



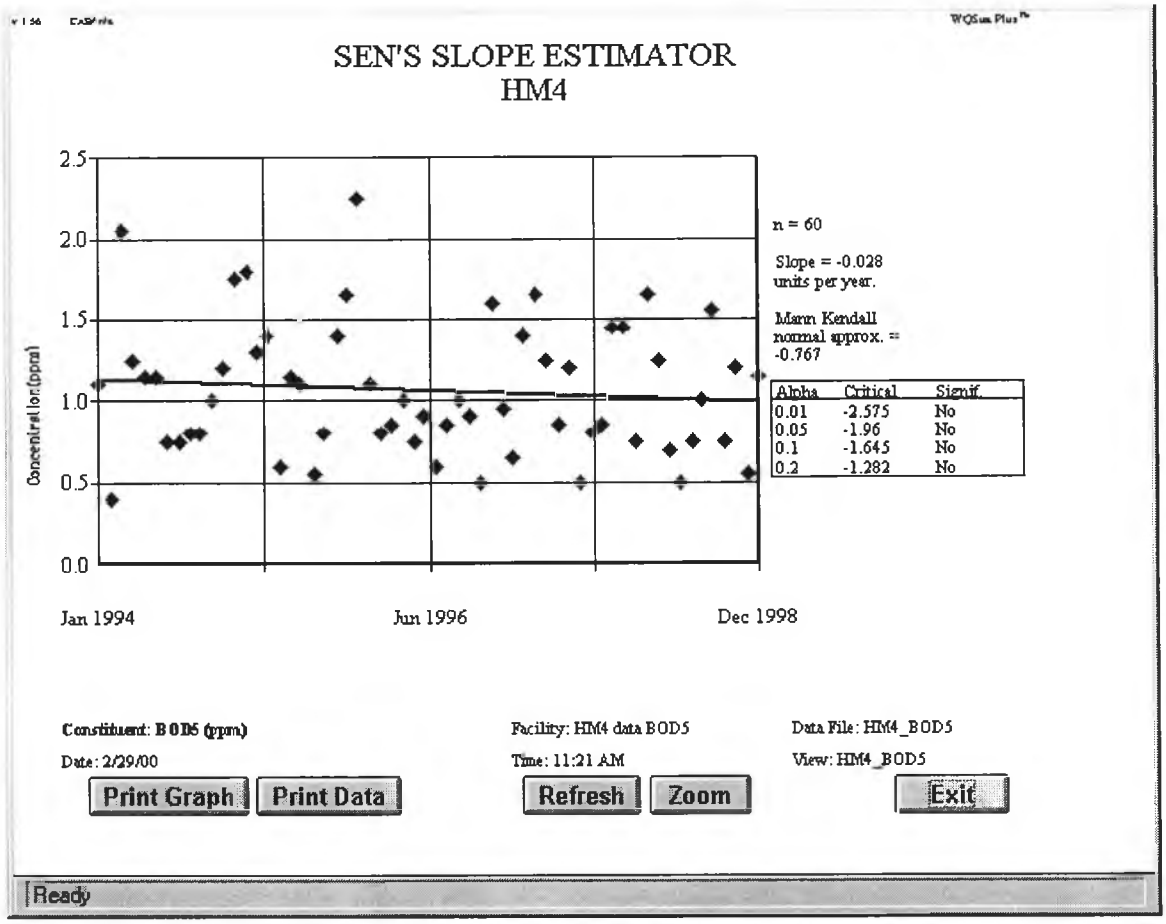
Appendix G3. Mann-Kendall Results for HM4_BOD5

1st 5-yr, FAC



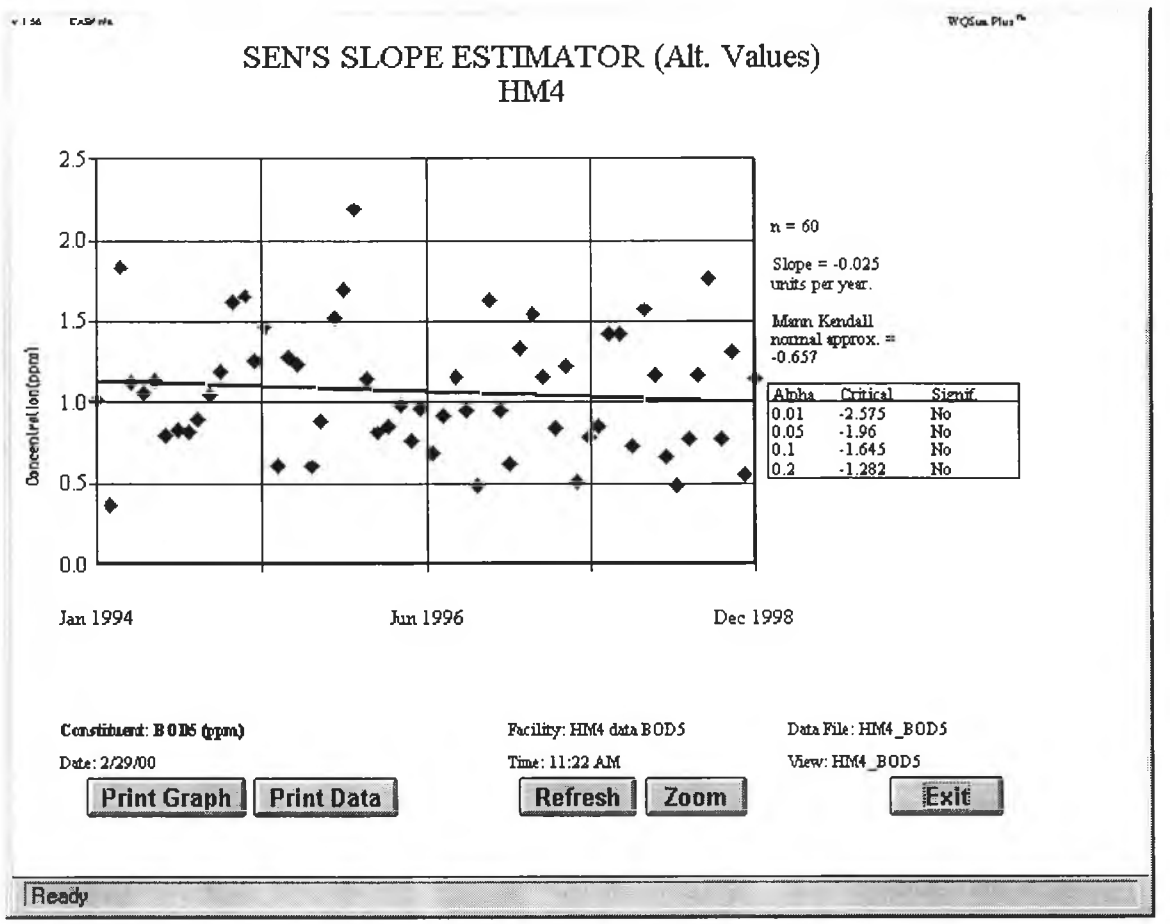
Appendix G3. Mann-Kendall Results for HM4_BOD5

2nd 5-yr, raw



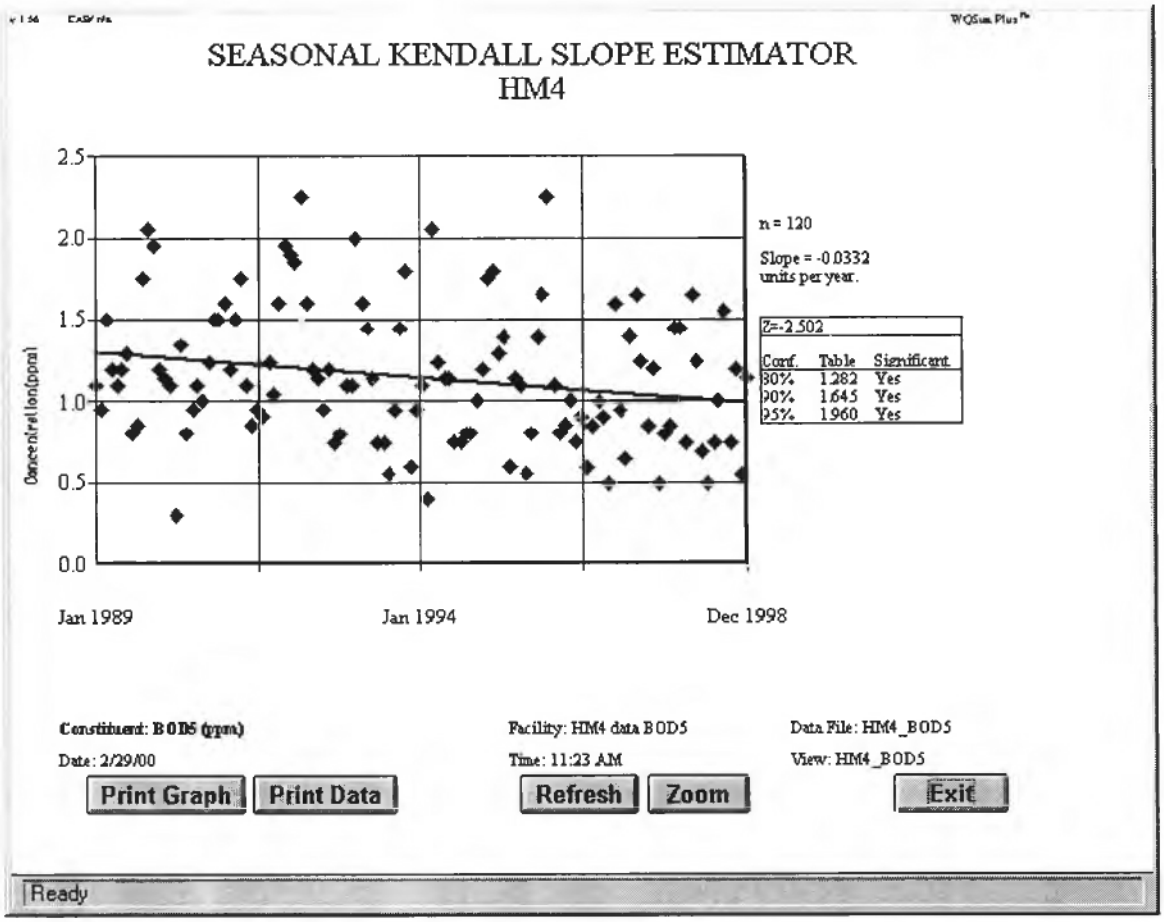
Appendix G3. Mann-Kendall Results for HM4_BOD5

2nd 5-yr, FAC



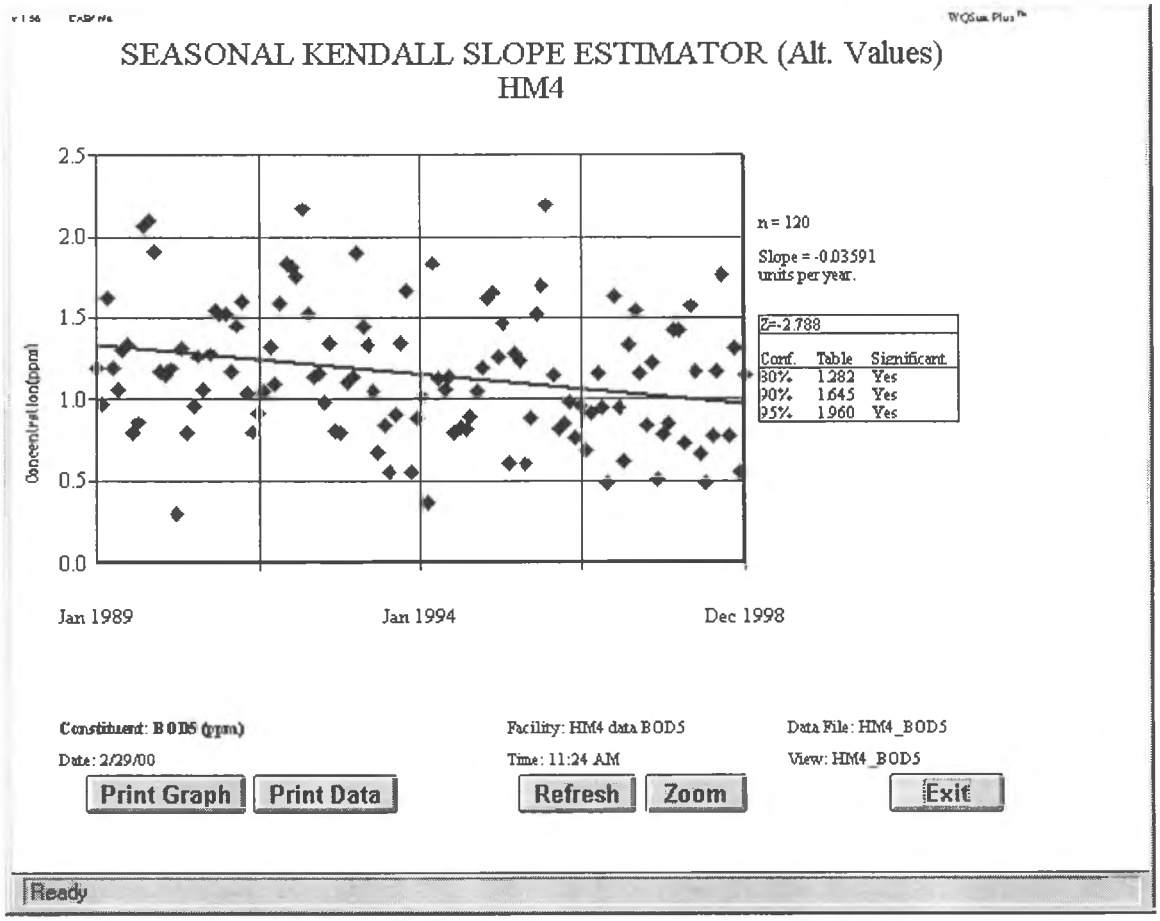
Appendix G4. Seasonal Kendall Results for HM4_BOD5

10-yr, raw



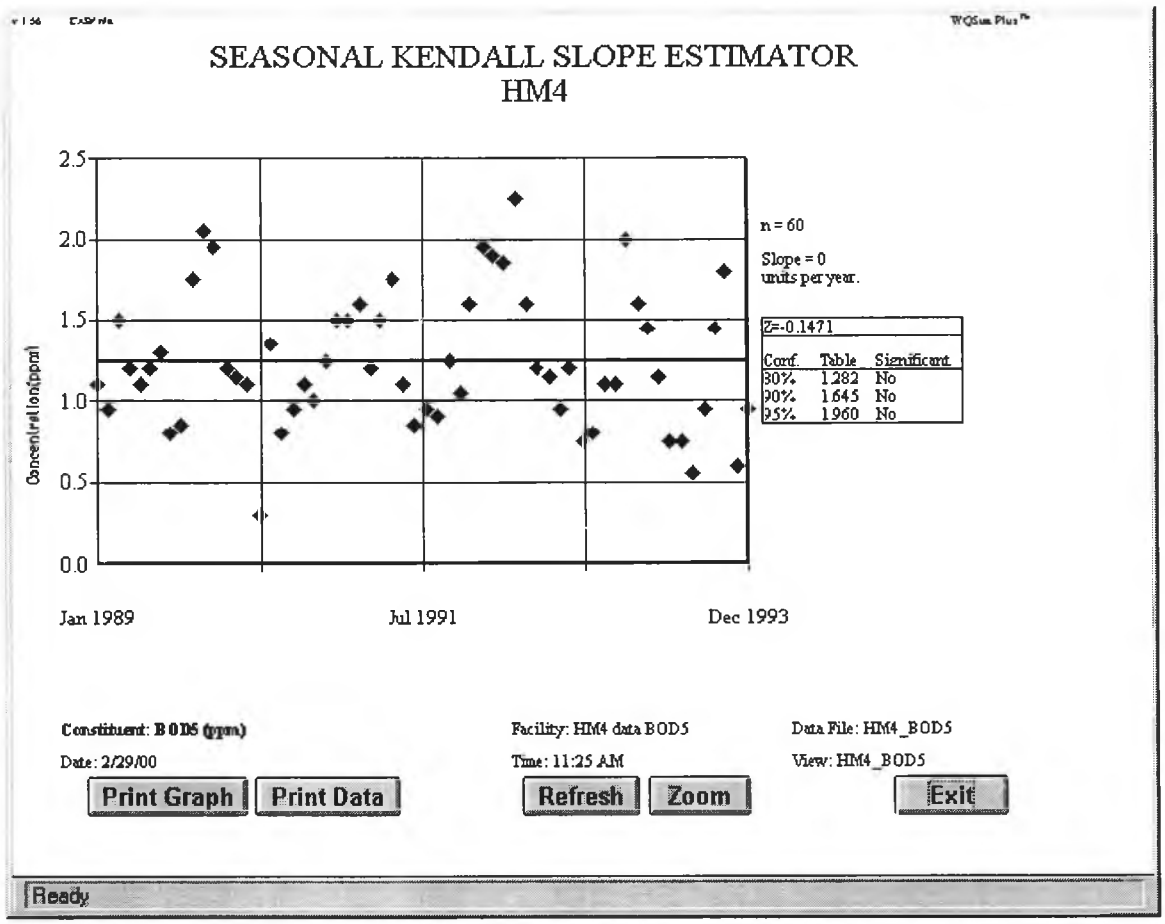
Appendix G4. Seasonal Kendall Results for HM4_BOD5

10-yr, FAC



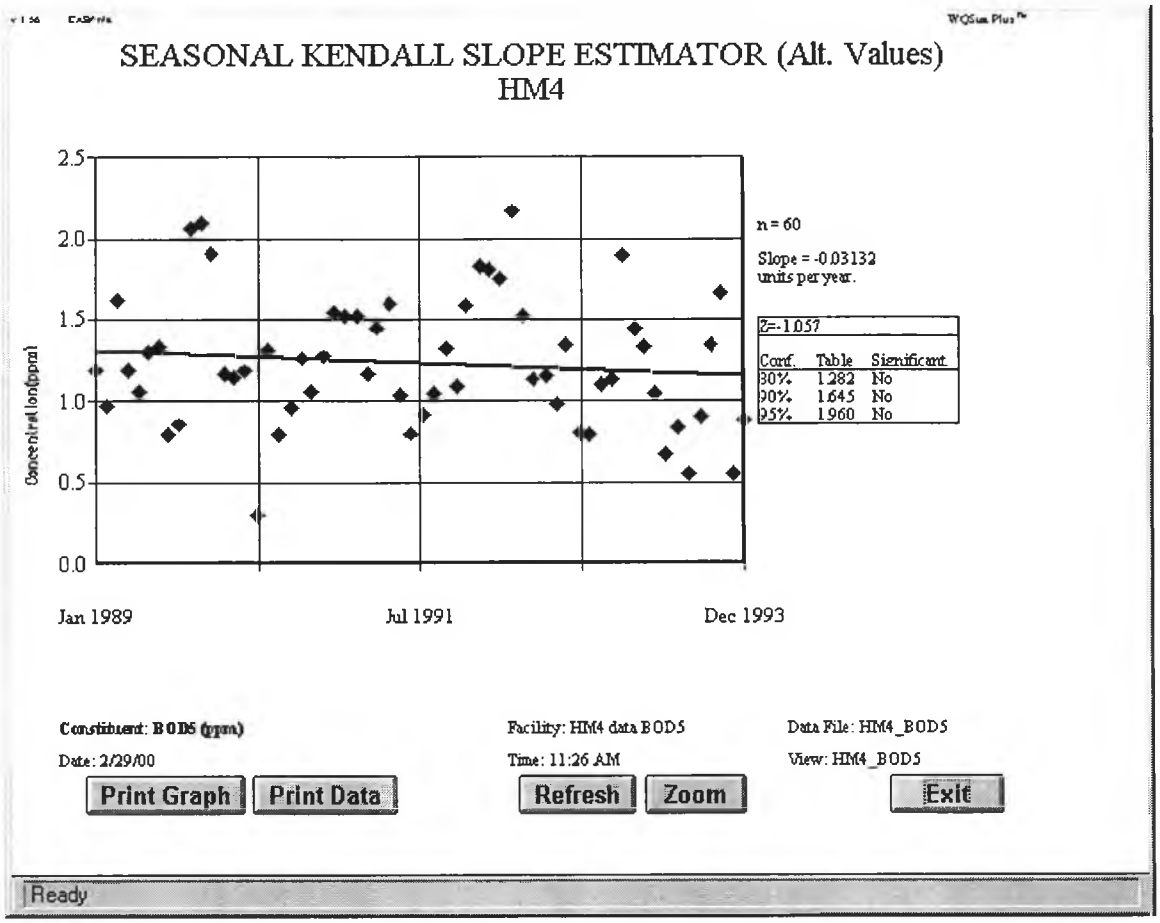
Appendix G4. Seasonal Kendall Results for HM4_BOD5

1st 5-yr, raw



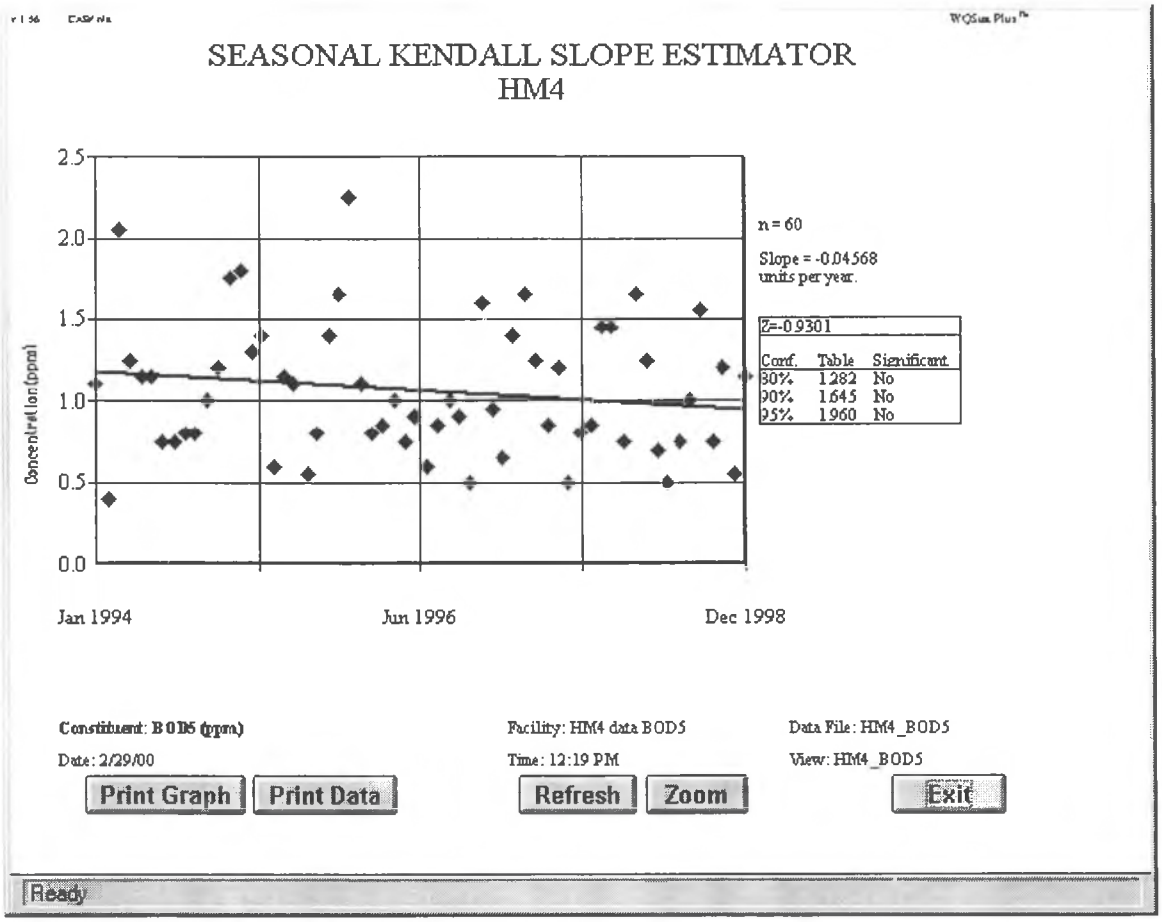
Appendix G4. Seasonal Kendall Results for HM4_BOD5

1st 5-yr, FAC



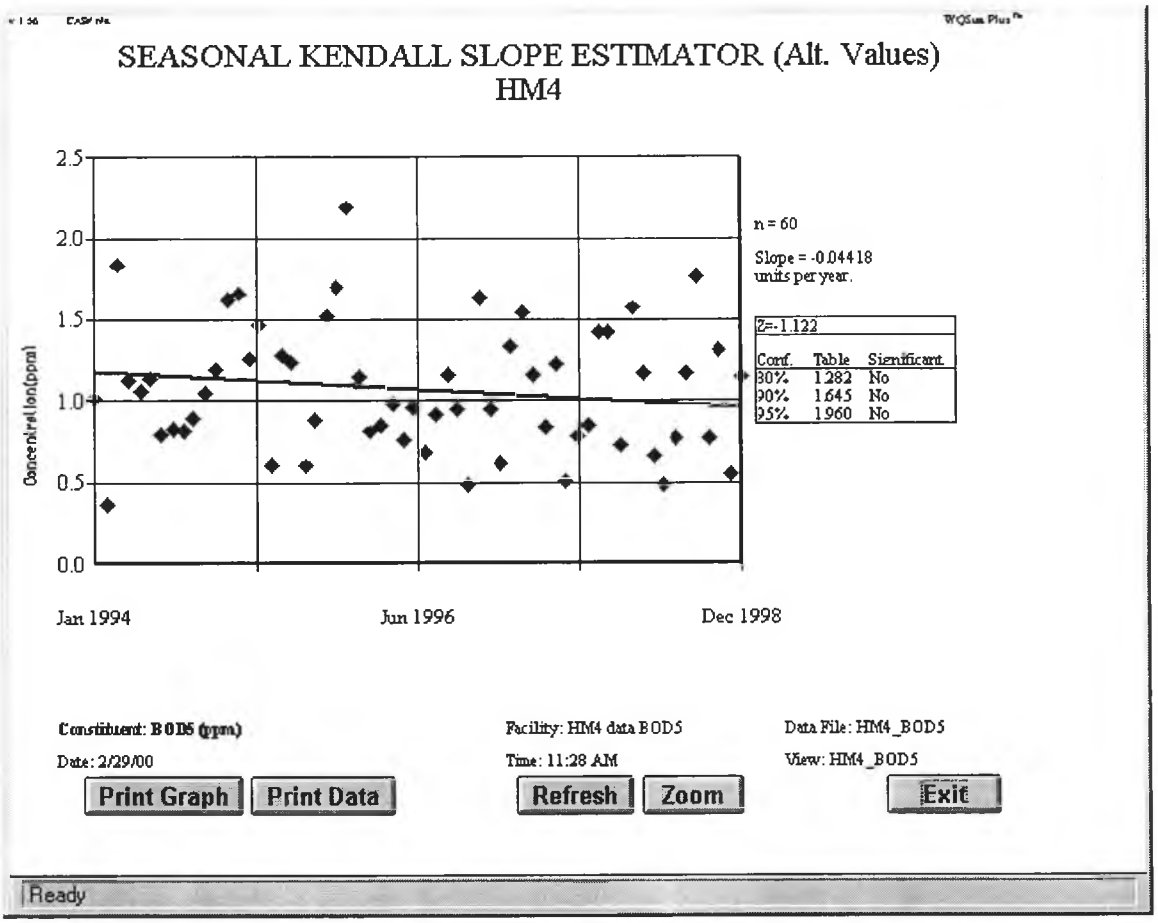
Appendix G4. Seasonal Kendall Results for HM4_BOD5

2nd 5-yr, raw



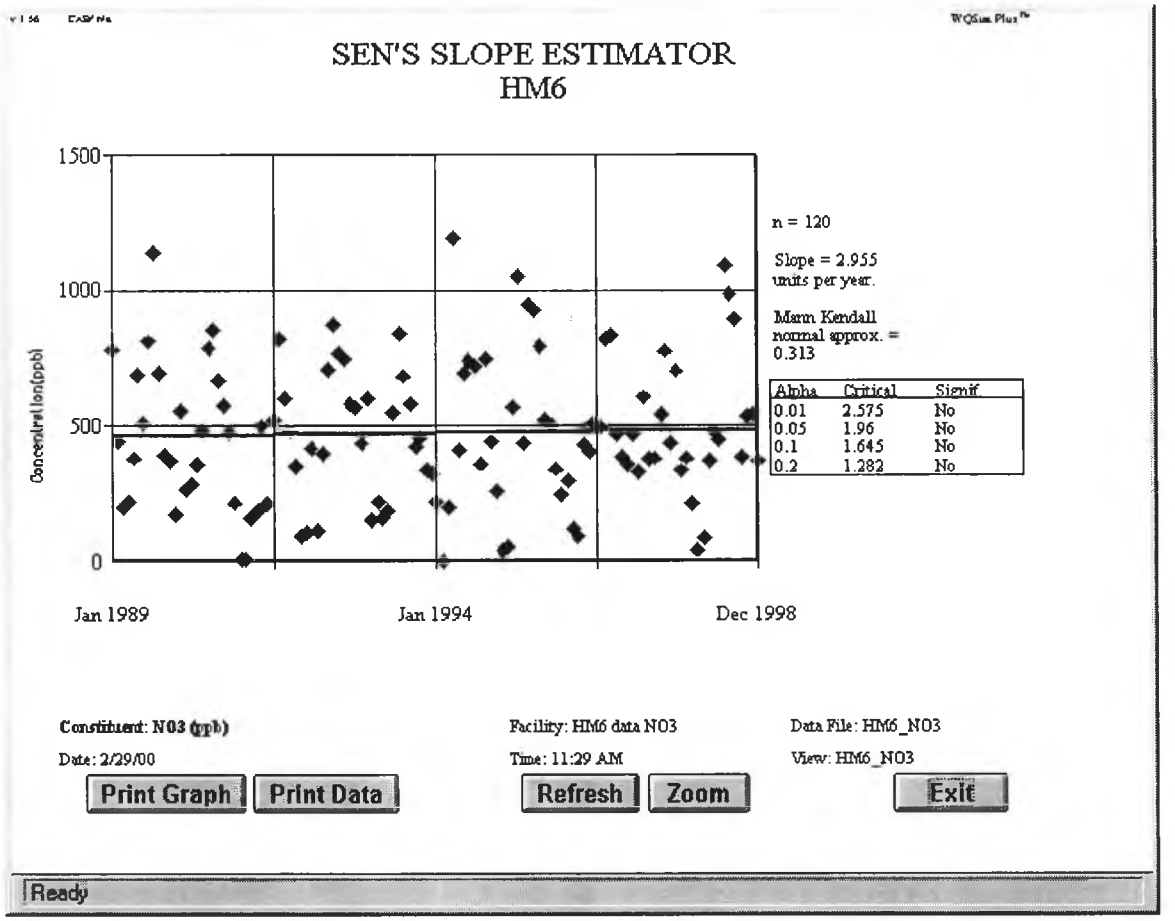
Appendix G4. Seasonal Kendall Results for HM4_BOD5

2nd 5-yr, FAC



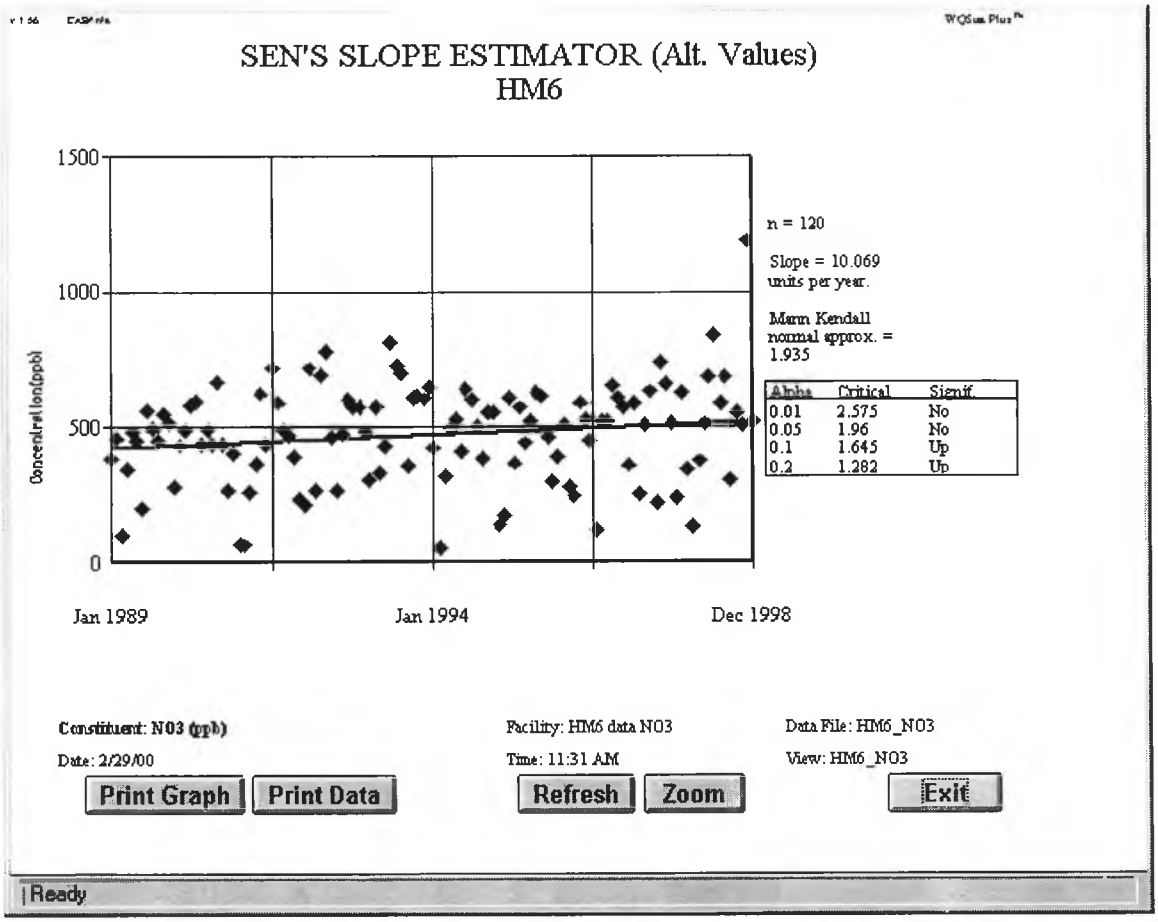
Appendix G5. Mann-Kendall Results for HM6_NO3

10-yr, raw



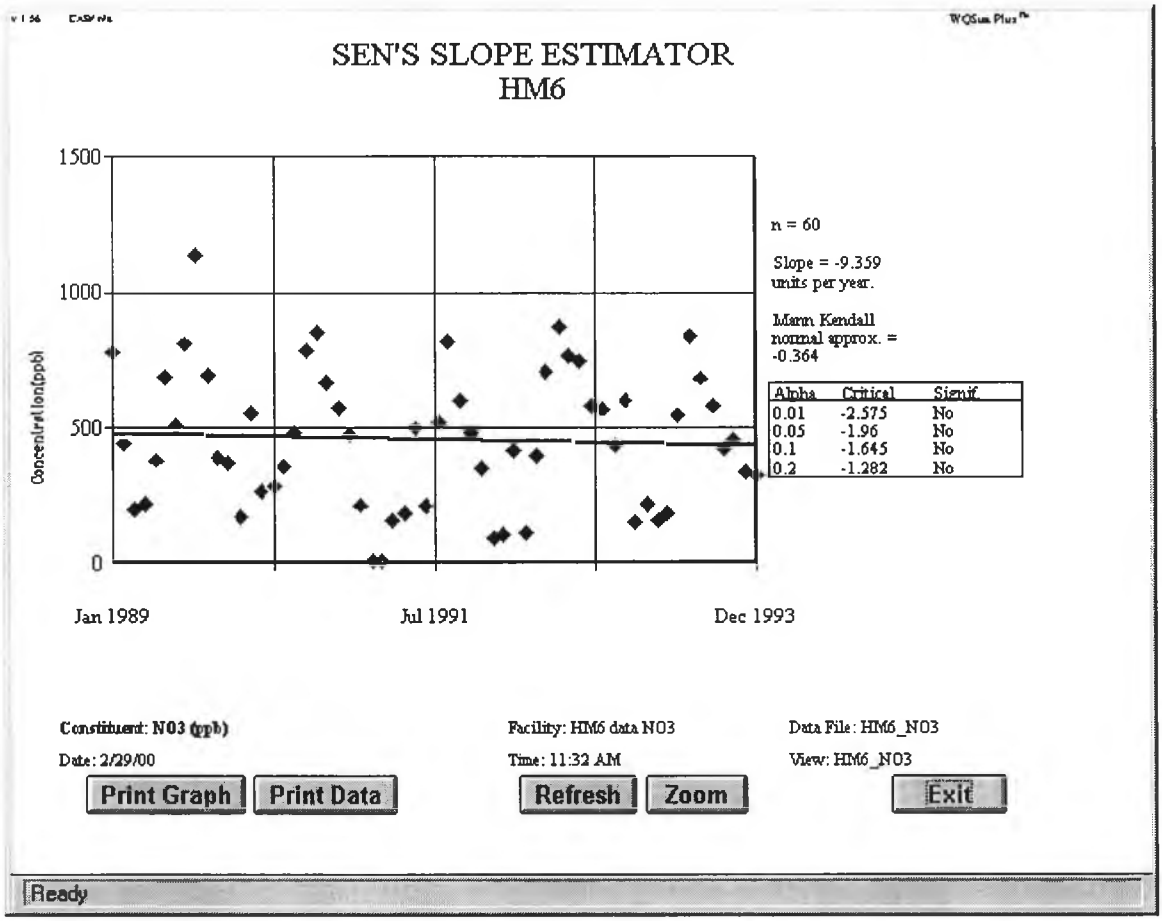
Appendix G5. Mann-Kendall Results for HM6_NO3

10-yr, FAC



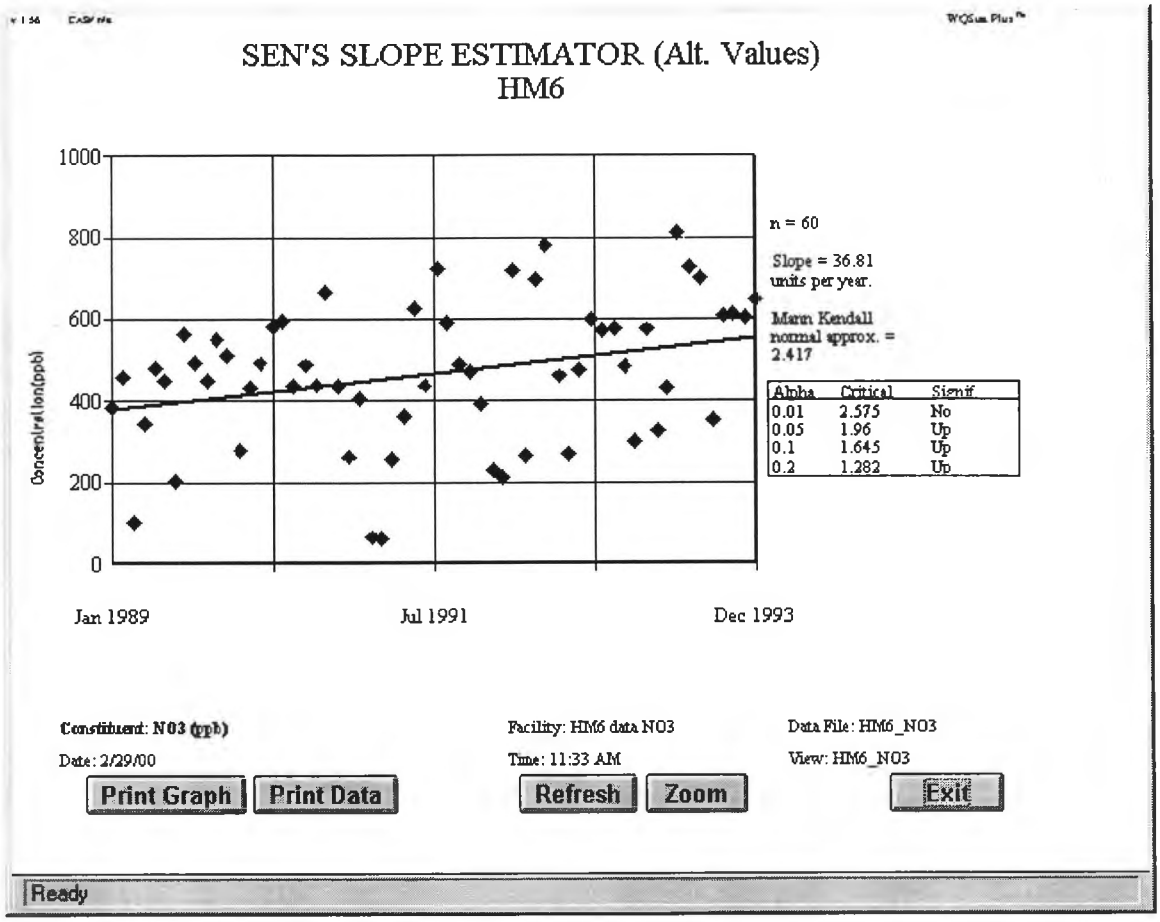
Appendix G5. Mann-Kendall Results for HM6_NO3

1st 5-yr, raw



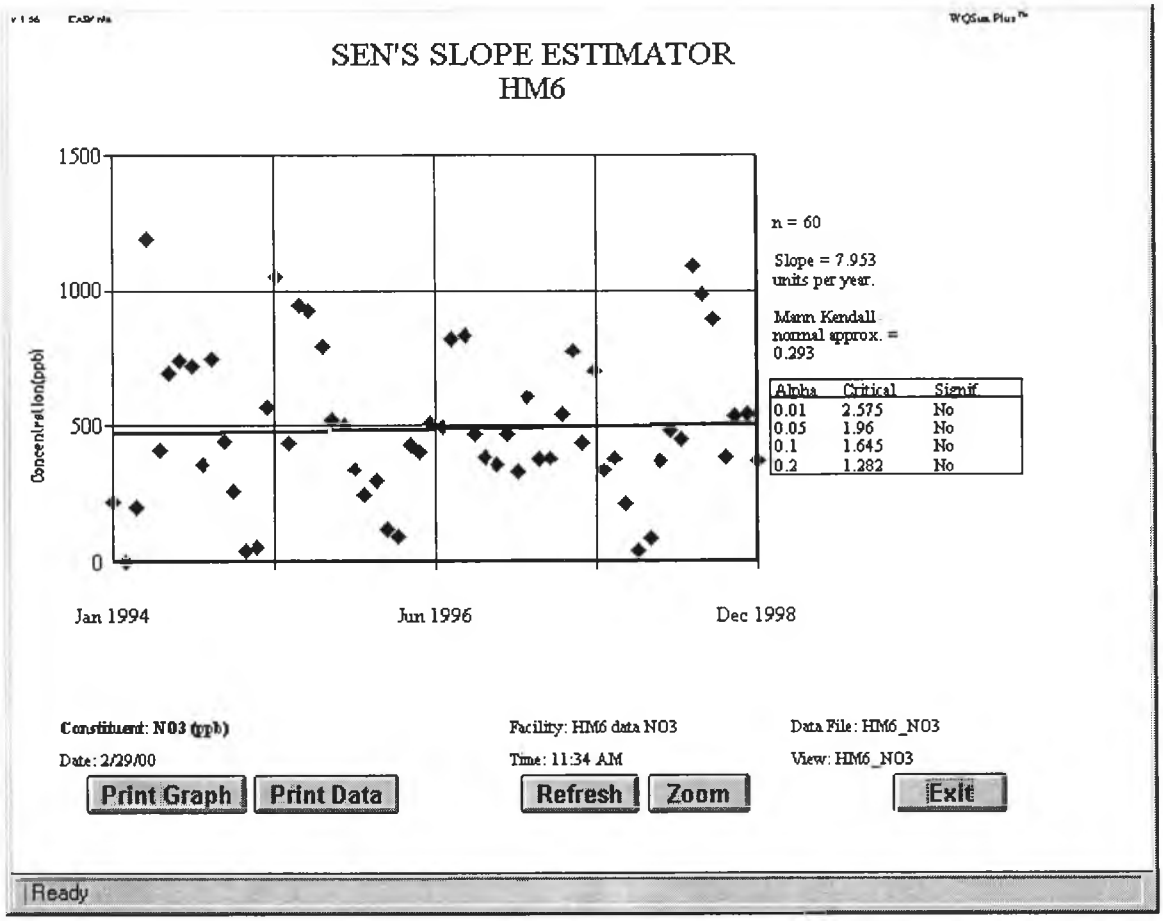
Appendix G5. Mann-Kendall Results for HM6_NO3

1st 5-yr, FAC



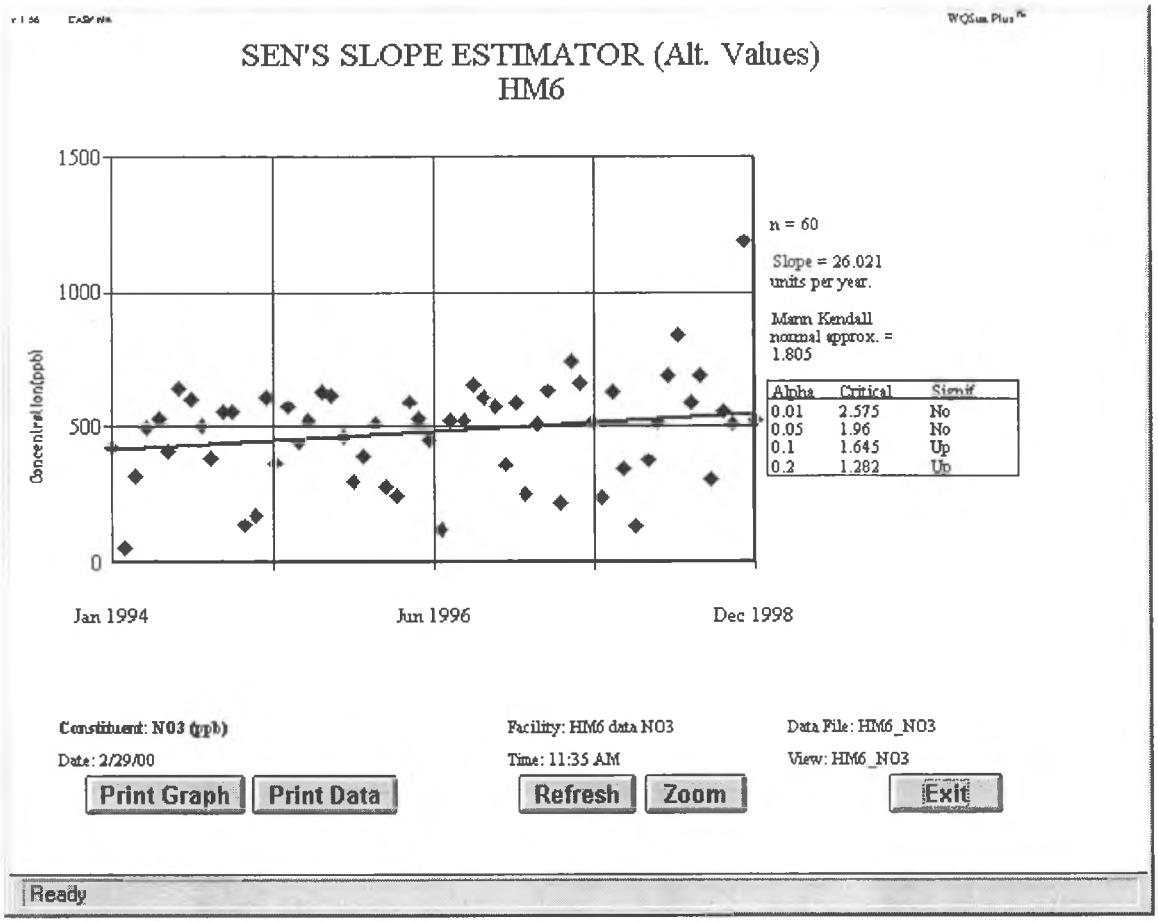
Appendix G5. Mann-Kendall Results for HM6_NO3

2nd 5-yr, raw



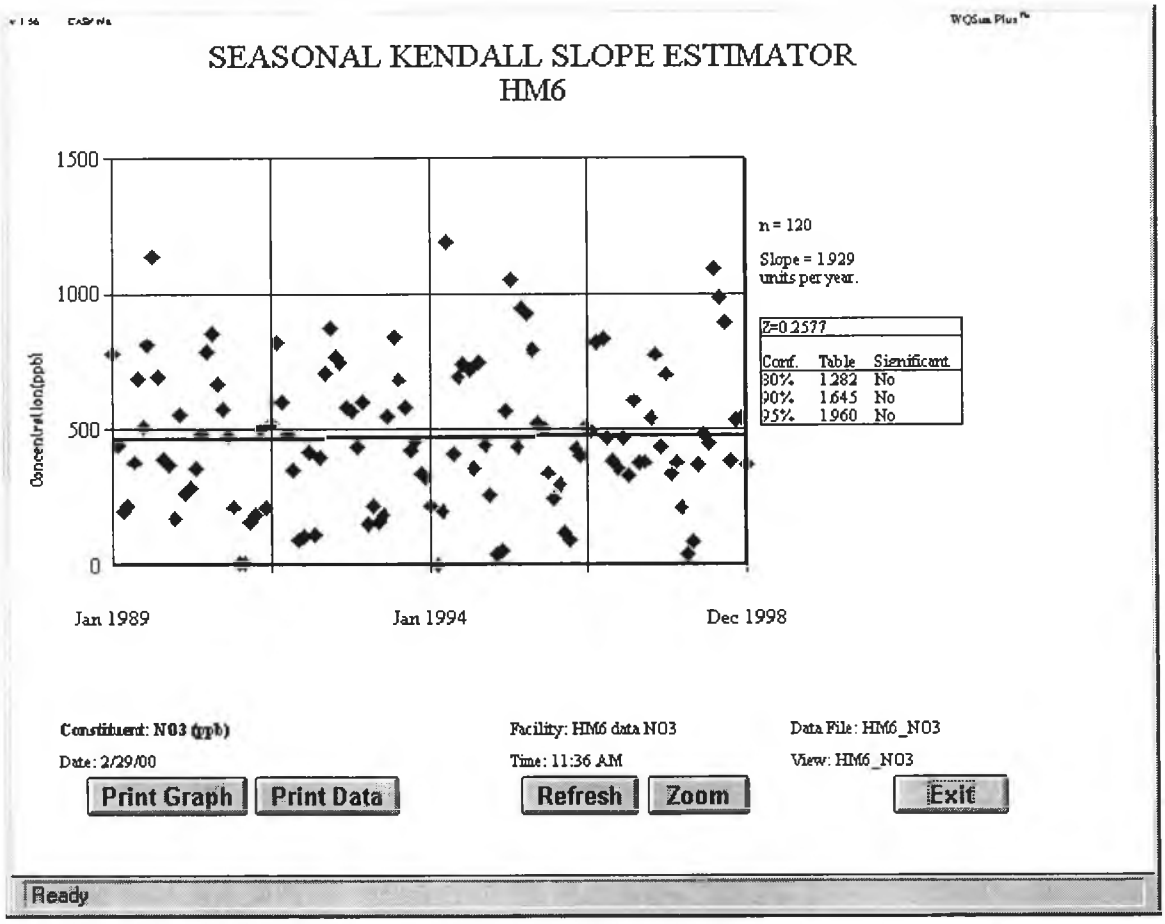
Appendix G5. Mann-Kendall Results for HM6_NO3

2nd 5-yr, FAC



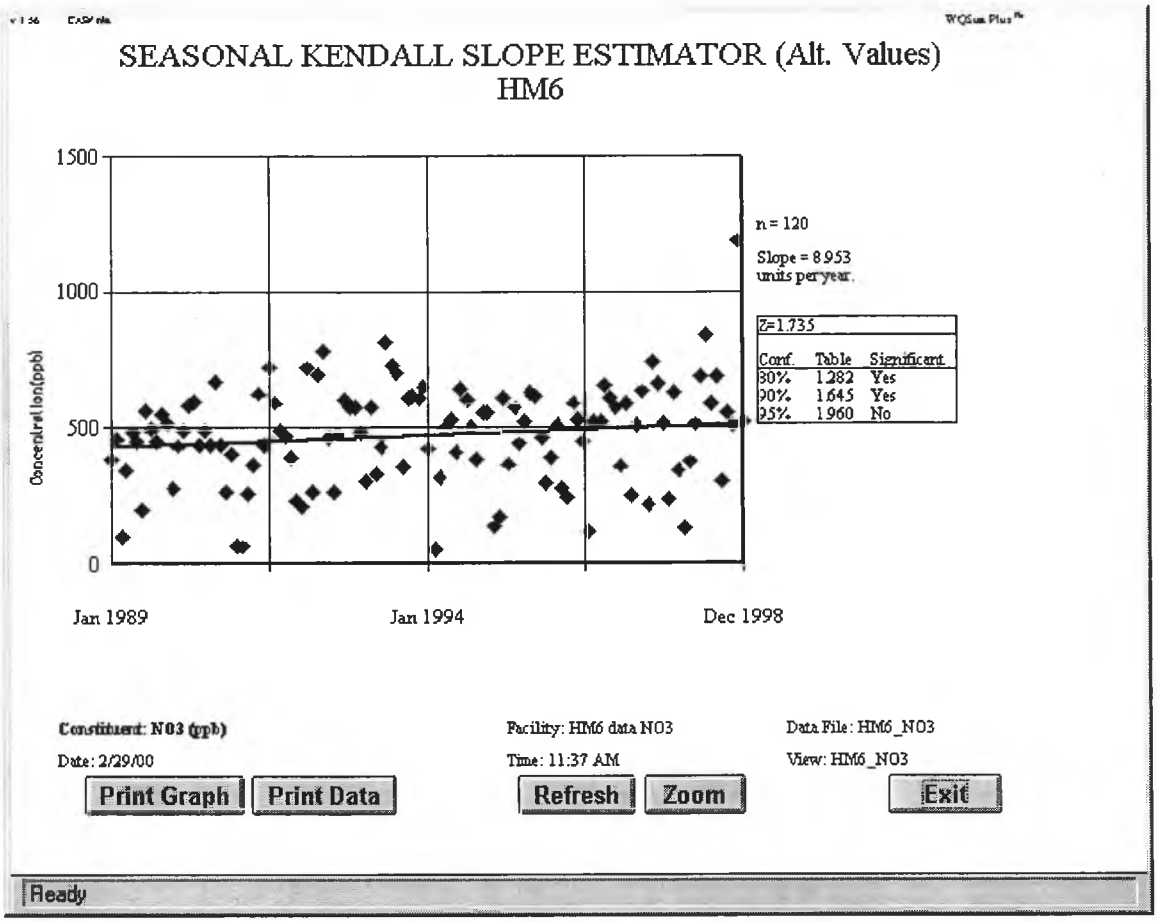
Appendix G6. Seasonal Kendall Results for HM6_NO3

10-yr, raw



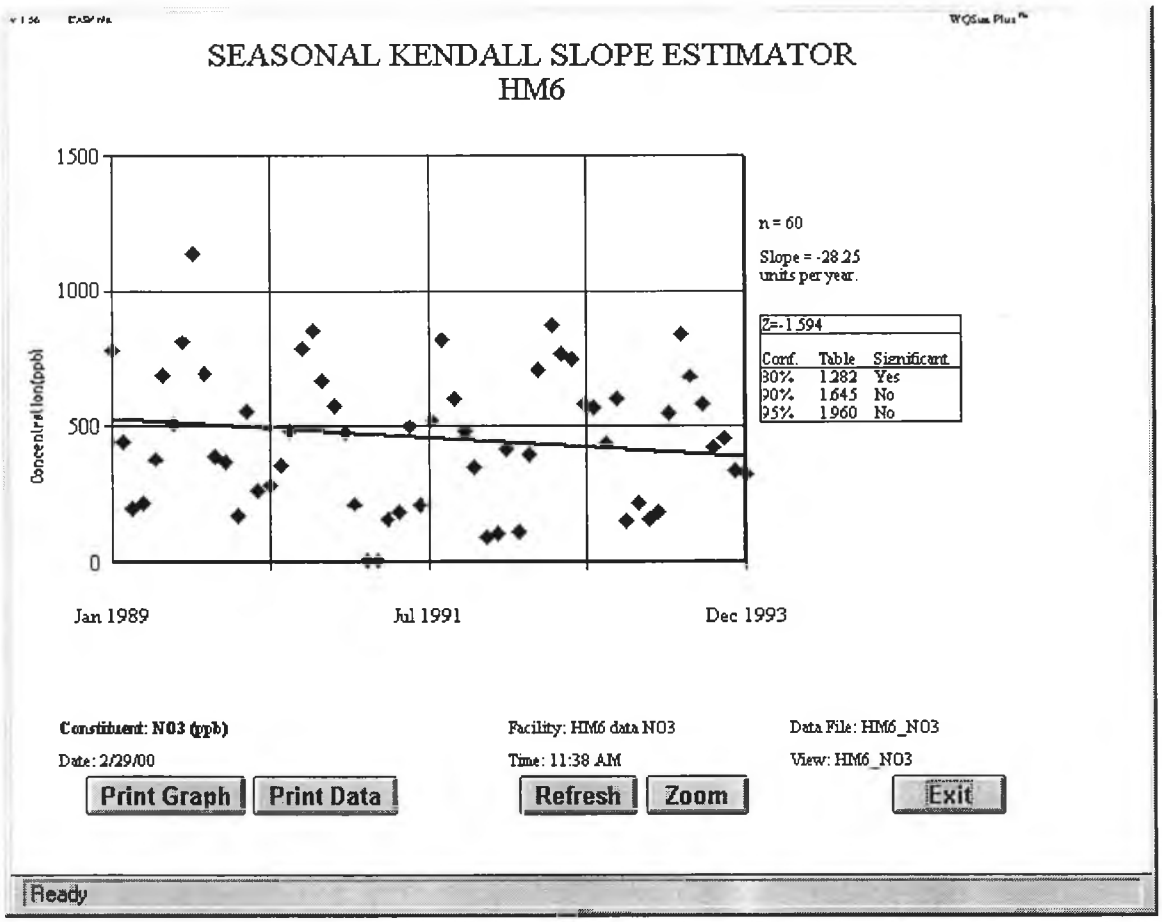
Appendix G6. Seasonal Kendall Results for HM6_NO3

10-yr, FAC



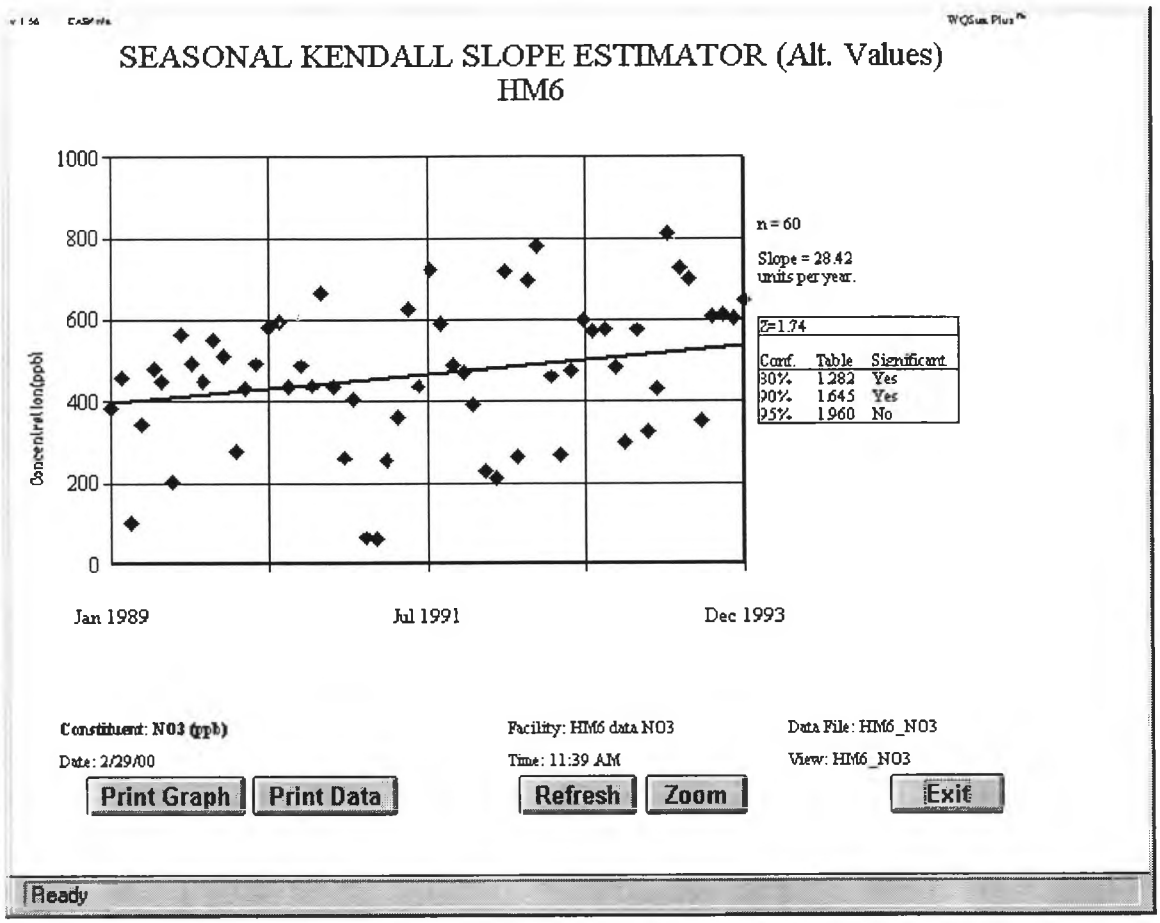
Appendix G6. Seasonal Kendall Results for HM6_NO3

1st 5-yr, raw



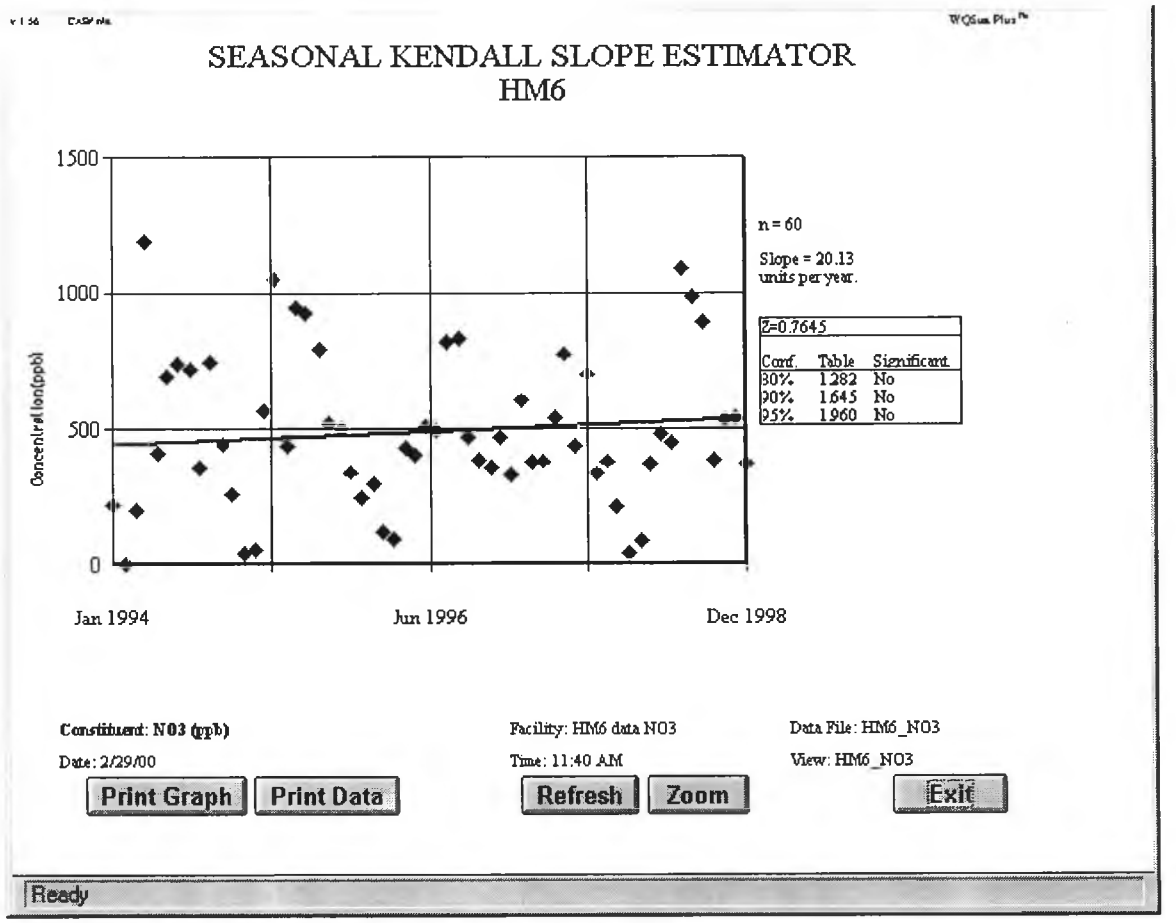
Appendix G6. Seasonal Kendall Results for HM6_NO3

1st 5-yr, FAC



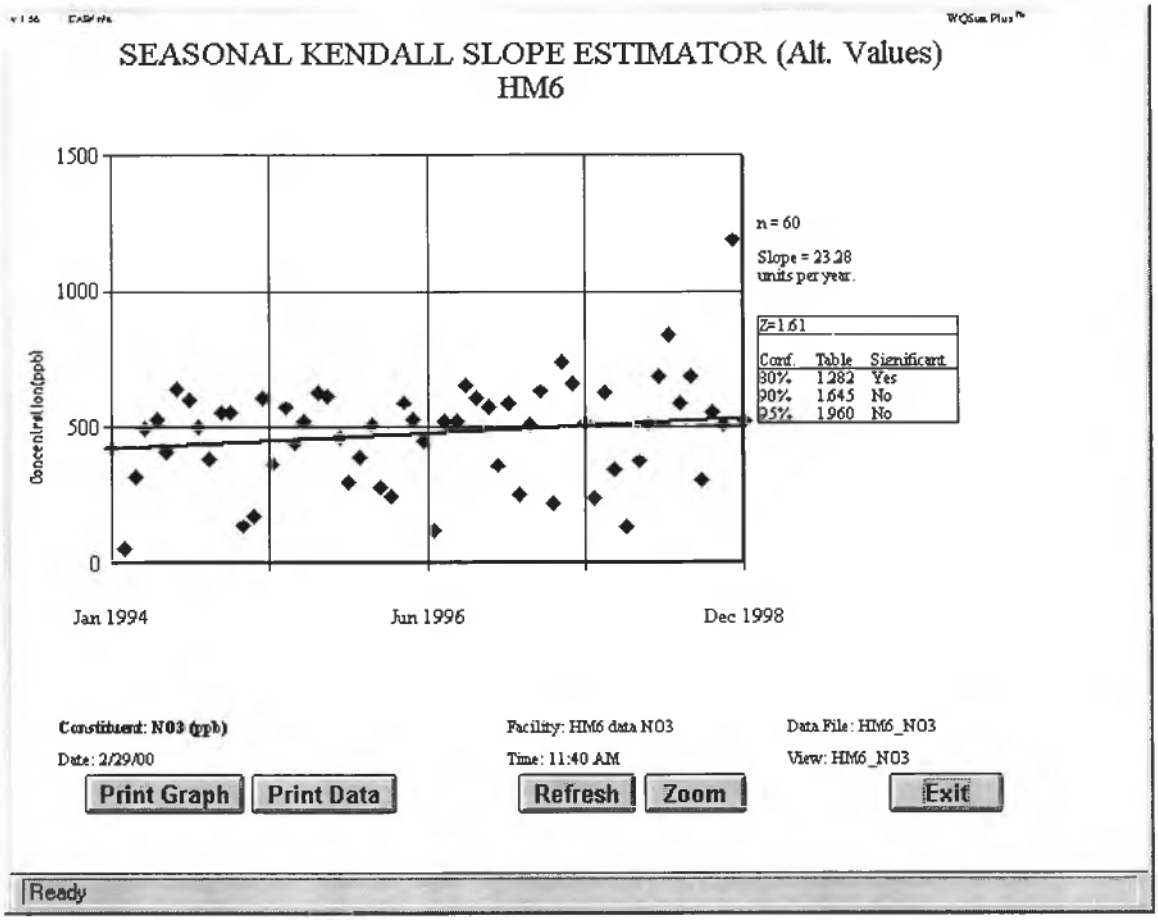
Appendix G6. Seasonal Kendall Results for HM6_NO3

2nd 5-yr, raw



Appendix G6. Seasonal Kendall Results for HM6_NO3

2nd 5-yr, FAC



Appendix H.
F-test for Equal Variances Results in MS-Excel™

Appendix H. F-test for equal variances results

F-Test Two-Sample for Variances
HM4_BOD5

	Variable 1	*Variable 2*
Mean	1.241667	1.063333
Variance	0.170438	0.164311
Observations	60	60
df	59	59
F	1.03729	
P(F<=f) one-tail	0.444326	
F Critical one-tail	1.539956	

F-Test Two-Sample for Variances
RO2_NH4

	Variable 1	*Variable 2*
Mean	48.52542	53.1875
Variance	362.2881	249.8577
Observations	59	48
df	58	47
F	1.449978	
P(F<=f) one-tail	0.095043	
F Critical one-tail	1.59554	

F-Test Two-Sample for Variances
HM6_NO3

	Variable 1	*Variable 2*
Mean	458.4833	490.3667
Variance	62948.15	77484.71
Observations	60	60
df	59	59
F	0.812394	
P(F<=f) one-tail	0.213636	
F Critical one-tail	0.649369	

F-Test Two-Sample for Variances
RO1&RO2

	Variable 1	*Variable 2*
Mean	3.320755	50.61682
Variance	5.019946	314.4461
Observations	106	107
df	105	106
F	0.015964	
P(F<=f) one-tail	0	
F Critical one-tail	0.724789	

Appendix I.
Differences in Populations Results in MS-Excel™, Minitab™, and WQStat Plus™

Appendix I1. T-test and Interval Test Results in MS-Excel™

Input data	BOD5	NO3	NH4	RO1&RO2
nup =	60	60	59	106
ndown=	60	60	48	107
a =	5	5	5	5
Dl =	-20	-20	-20	-20
Du =	20	20	20	20
df =	118	118	105	211
t(a2, df)	1.98027	1.98027	1.982817	1.9712706
t(a,df)	1.65787	1.65787	1.659496	1.6521062
Derived data				
xbar-up	1.241667	458.4833	48.52542	3.3207547
xbar-down	1.063333	490.3667	53.1875	50.616822
delta	-0.17833	31.88333	4.662076	47.296068
sup	0.412841	250.8947	19.03387	2.2405236
sdown	0.405353	278.3608	15.80689	17.732629
spooled	0.409114	264.9838	17.65933	12.667257
SE	0.074694	48.37921	3.432571	1.7359114
dl	-0.24833	-91.6967	-9.70508	-0.664151
du	0.248333	91.69667	9.705085	0.6641509
Tl	0.937161	2.554403	4.185539	27.628265
Tu	-5.71222	-1.23634	-1.46916	26.863075
T	2.38753	0.65903	1.358188	27.24567
Results				
null (t-test)	rejected	accept	accept	reject
equivalence	accepted	accept	accept	reject
inequivalence	accepted	rejected	rejected	accept

Appendix I2. T-test Results in Minitab™

Two Sample T-Test and Confidence Interval for HM4_BOD5

```
Two sample T for BOD5 (ppm) vs BOD5

           N      Mean     StDev   SE Mean
BOD5 (pp  60     1.242     0.413     0.053
BOD5      60     1.063     0.405     0.052

95% CI for mu BOD5 (pp - mu BOD5: ( 0.030,  0.326)
T-Test mu BOD5 (pp = mu BOD5 (vs not =): T = 2.39  P = 0.019  DF = 118
Both use Pooled StDev = 0.409

Saving file as: C:\USERS\Lindsay\thesis\BOD5.MTW
```

Two Sample T-Test and Confidence Interval for RO2_NH4 vs. RO1_NH4

```
Two sample T for NH4 vs NH4(2)

           N      Mean     StDev   SE Mean
NH4      106      3.32      2.24      0.22
NH4(2)   107      50.6      17.7       1.7

95% CI for mu NH4 - mu NH4(2): ( -50.72,  -43.9)
T-Test mu NH4 = mu NH4(2) (vs not =): T = -27.25  P = 0.0000  DF = 211
Both use Pooled StDev = 12.7
```

Two Sample T-Test and Confidence Interval for HM6_NO3

```
Two sample T for NO3 vs NO3 (2)

           N      Mean     StDev   SE Mean
NO3       60       458       251        32
NO3 (2)   60       490       278        36

95% CI for mu NO3 - mu NO3 (2): ( -128,  64)
T-Test mu NO3 = mu NO3 (2) (vs not =): T = -0.66  P = 0.51  DF = 118
Both use Pooled StDev =  265

Saving file as: C:\USERS\Lindsay\thesis\NO3.MTW
```

Two Sample T-Test and Confidence Interval for RO2_NH4

```
Two sample T for NH4 vs NH4 (2)

           N      Mean     StDev   SE Mean
NH4       59      48.5      19.0       2.5
NH4 (2)   48      53.2      15.8       2.3

95% CI for mu NH4 - mu NH4 (2): ( -11.5,  2.1)
T-Test mu NH4 = mu NH4 (2) (vs not =): T = -1.36  P = 0.18  DF = 105
Both use Pooled StDev = 17.7

Saving file as: C:\USERS\Lindsay\thesis\NH4.MTW
```

Appendix I3. Mann-Whitney Results in Minitab™

Mann-Whitney Confidence Interval and Test for RO1_NH4 vs. RO2_NH4

```
NH4         N = 106     Median =       3.000
NH4(2)      N = 107     Median =      52.000
Point estimate for ETA1-ETA2 is     -49.000
95.0 Percent CI for ETA1-ETA2 is (-51.999,-45.000)
W = 5696.0
Test of ETA1 = ETA2  vs  ETA1 not = ETA2 is significant at 0.0000
The test is significant at 0.0000 (adjusted for ties)
```

Mann-Whitney Confidence Interval and Test for HM4_BOD5

```
BOD5 (pp    N =  60     Median =      1.1750
BOD5        N =  60     Median =      1.0000
Point estimate for ETA1-ETA2 is       0.2000
95.0 Percent CI for ETA1-ETA2 is (0.0501,0.3499)
W = 4095.0
Test of ETA1 = ETA2  vs  ETA1 not = ETA2 is significant at 0.0148
The test is significant at 0.0146 (adjusted for ties)

Saving file as: C:\USERS\Lindsay\thesis\RO1_RO2.MTW
```

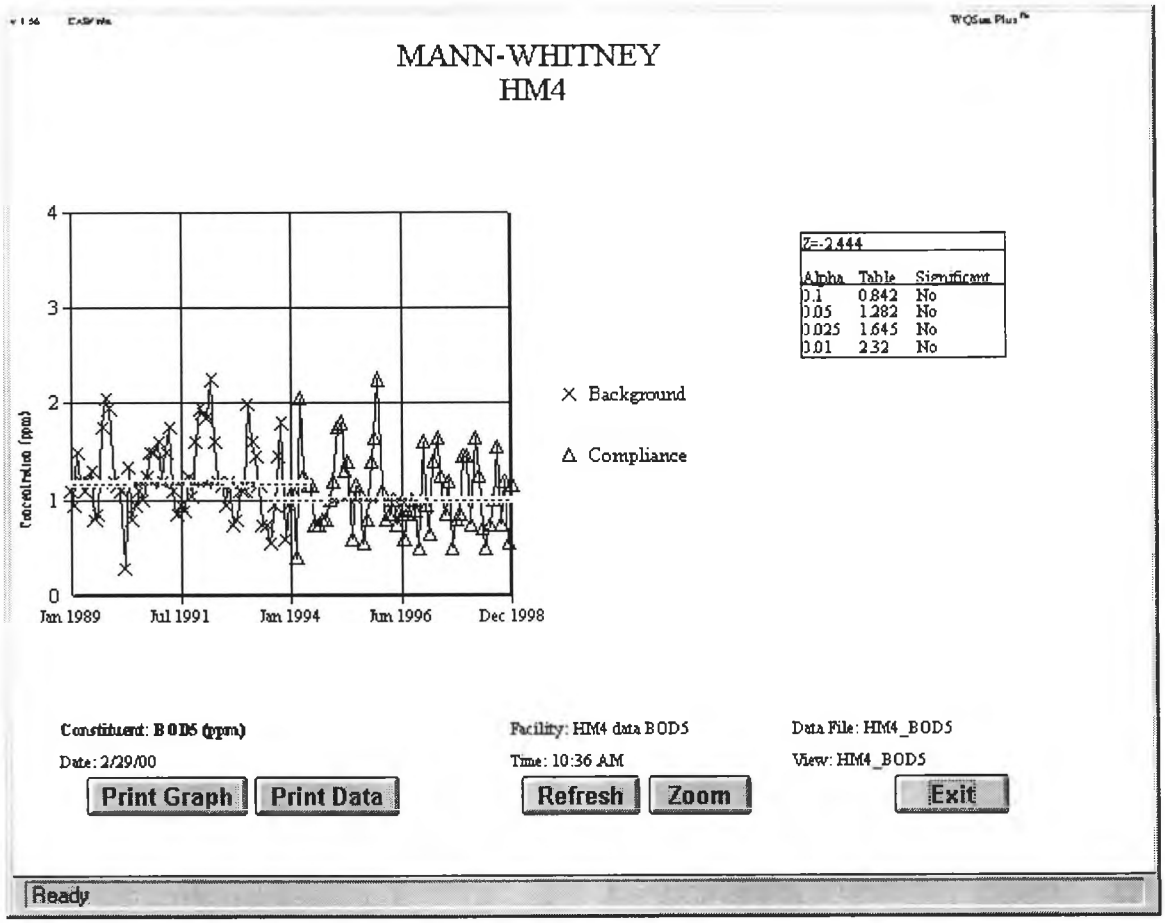
Mann-Whitney Confidence Interval and Test for RO2_NH4

```
NH4         N =  59     Median =       51.00
NH4 (2)     N =  48     Median =       53.50
Point estimate for ETA1-ETA2 is       -4.00
95.0 Percent CI for ETA1-ETA2 is (-10.99,3.00)
W = 3005.5
Test of ETA1 = ETA2  vs  ETA1 not = ETA2 is significant at 0.2595
The test is significant at 0.2594 (adjusted for ties)

Cannot reject at alpha = 0.05
```

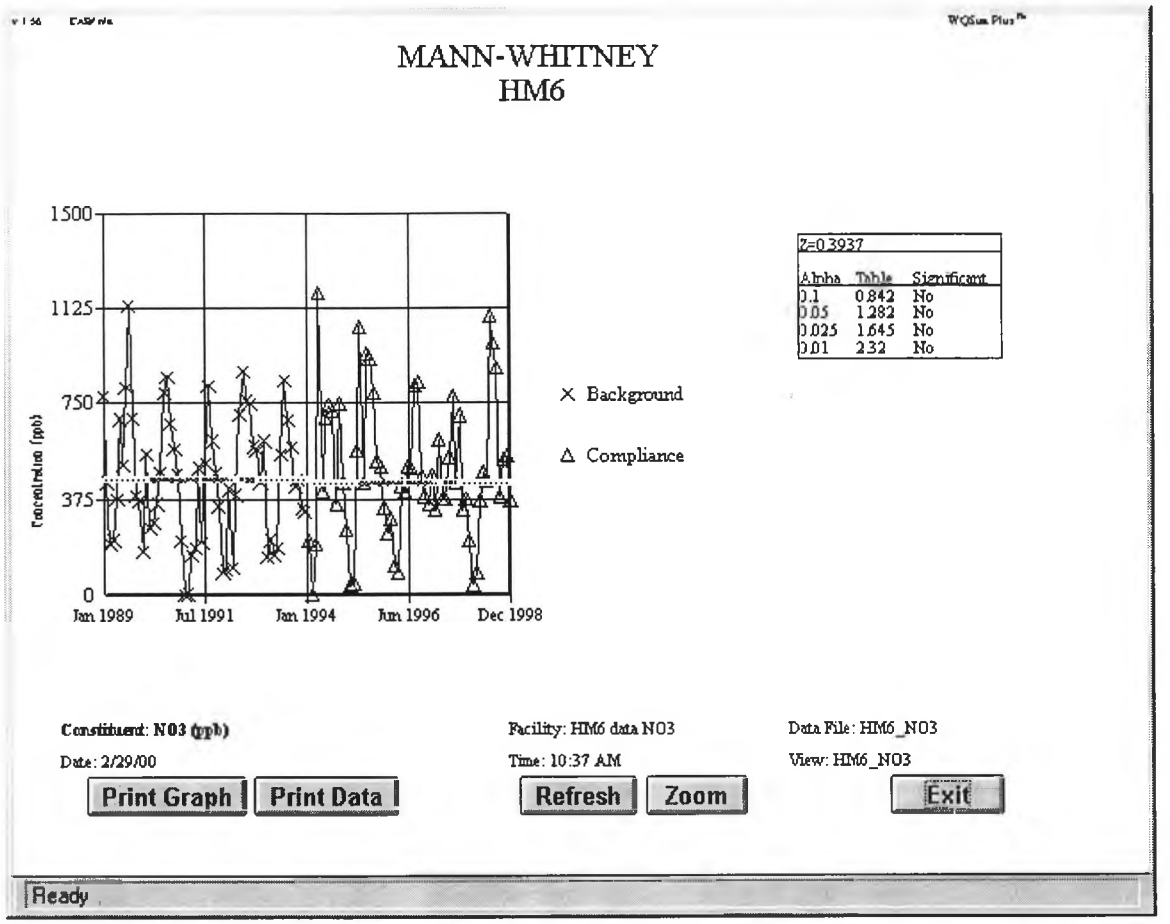
Mann-Whitney Confidence Interval and Test for HM6_NO3

```
NO3         N =  60     Median =       450.0
NO3 (2)     N =  60     Median =       441.0
Point estimate for ETA1-ETA2 is       -20.0
95.0 Percent CI for ETA1-ETA2 is (-126.0,75.0)
W = 3555.0
Test of ETA1 = ETA2  vs  ETA1 not = ETA2 is significant at 0.6958
The test is significant at 0.6958 (adjusted for ties)

Cannot reject at alpha = 0.05
```

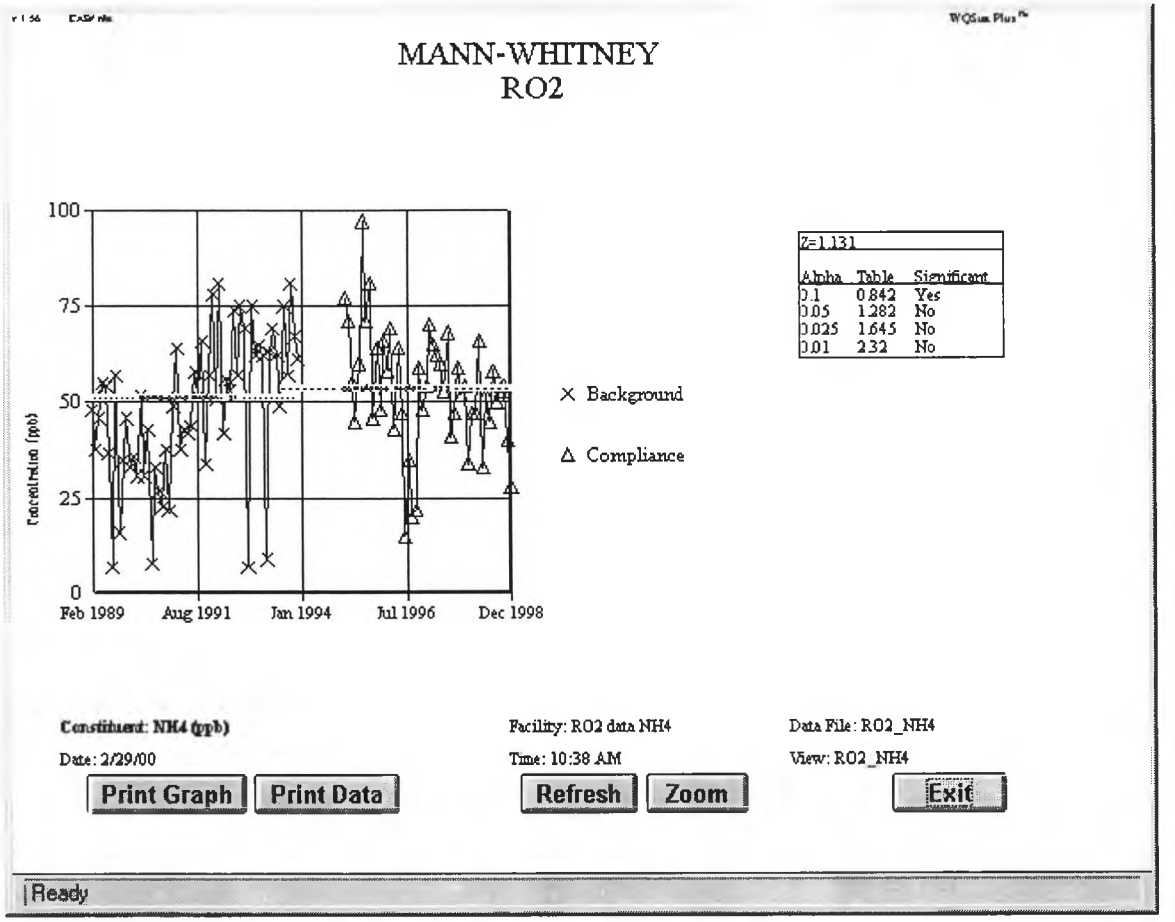
Appendix I4. Mann-Whitney Results in WQStat Plus™



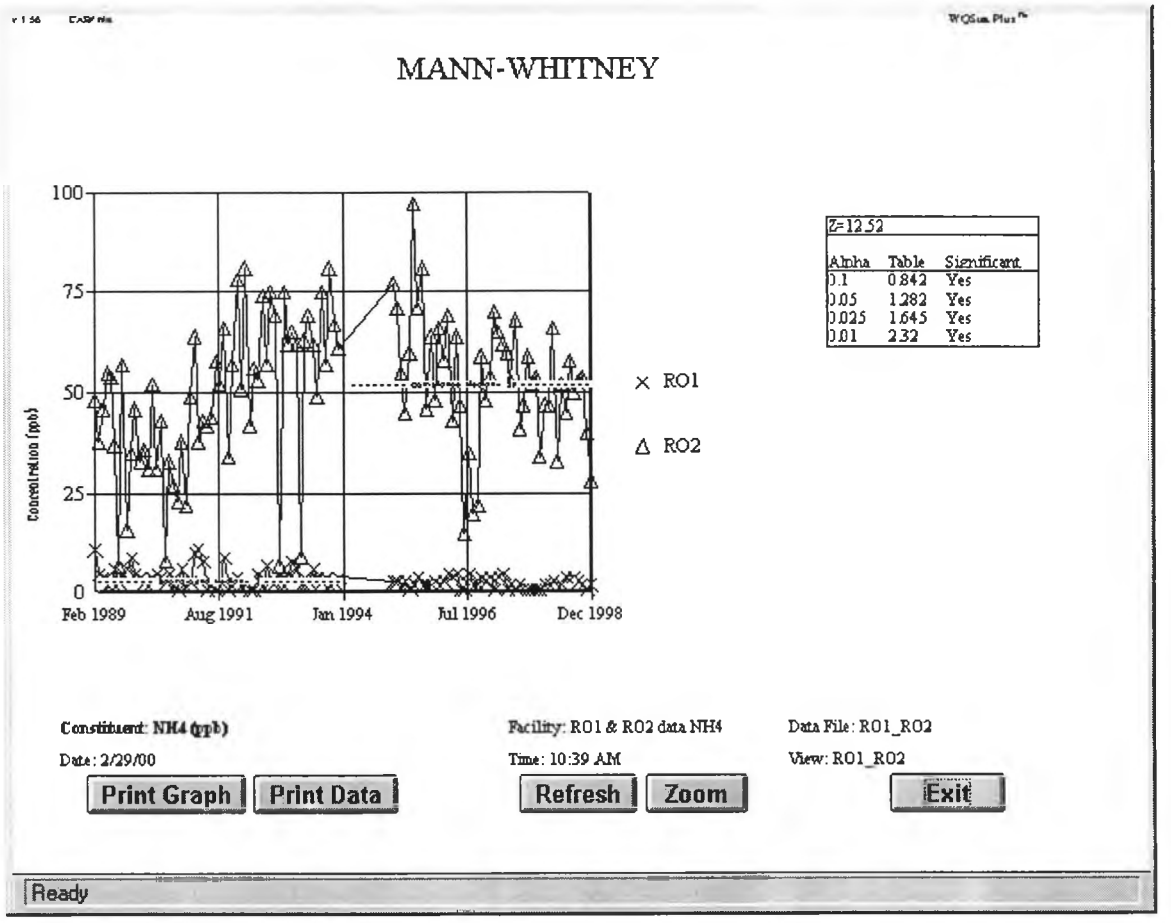
Appendix I4. Mann-Whitney Results in WQStat Plus™



Appendix I4. Mann-Whitney Results in WQStat Plus™



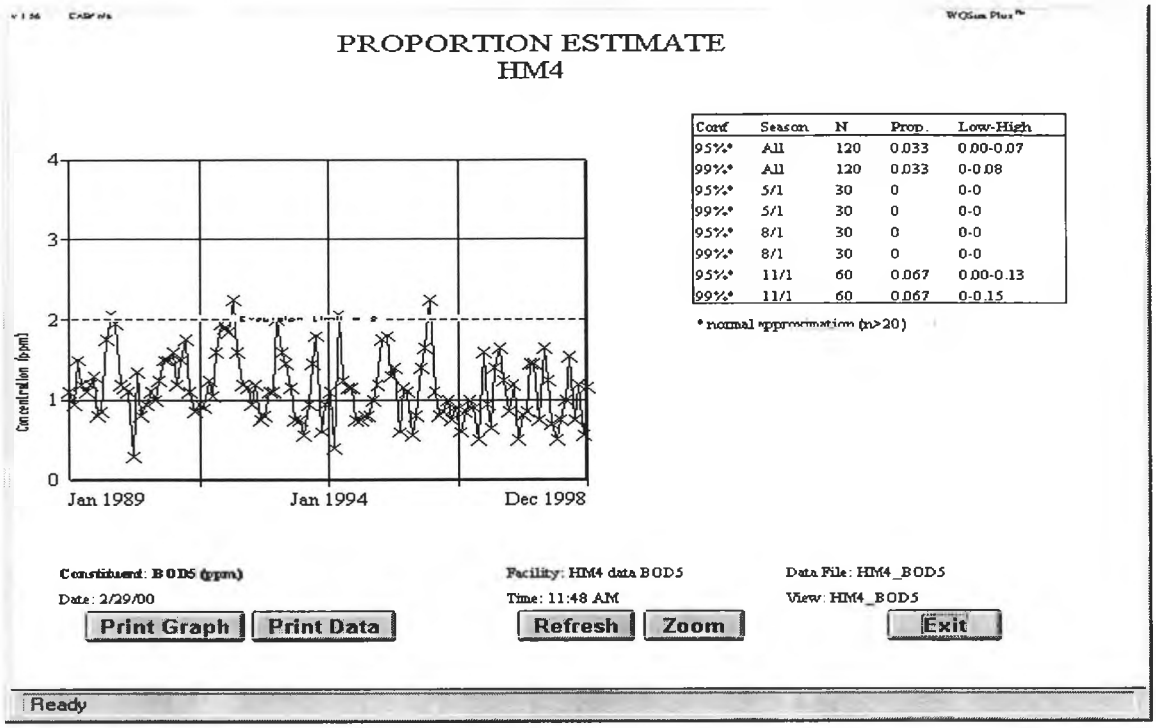
Appendix I4. Mann-Whitney Results in WQStat Plus™



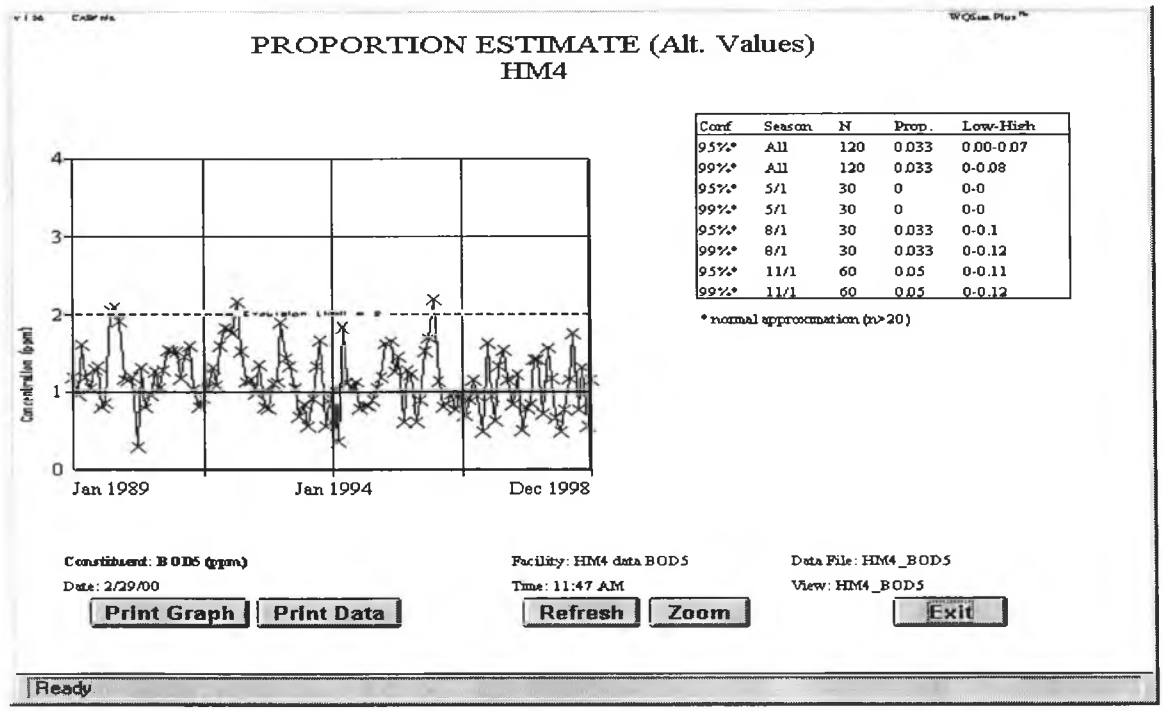
Appendix J.
Standards Compliance Results in WQStat Plus™

Appendix J. Standards Compliance for HM4_BOD5

Raw data

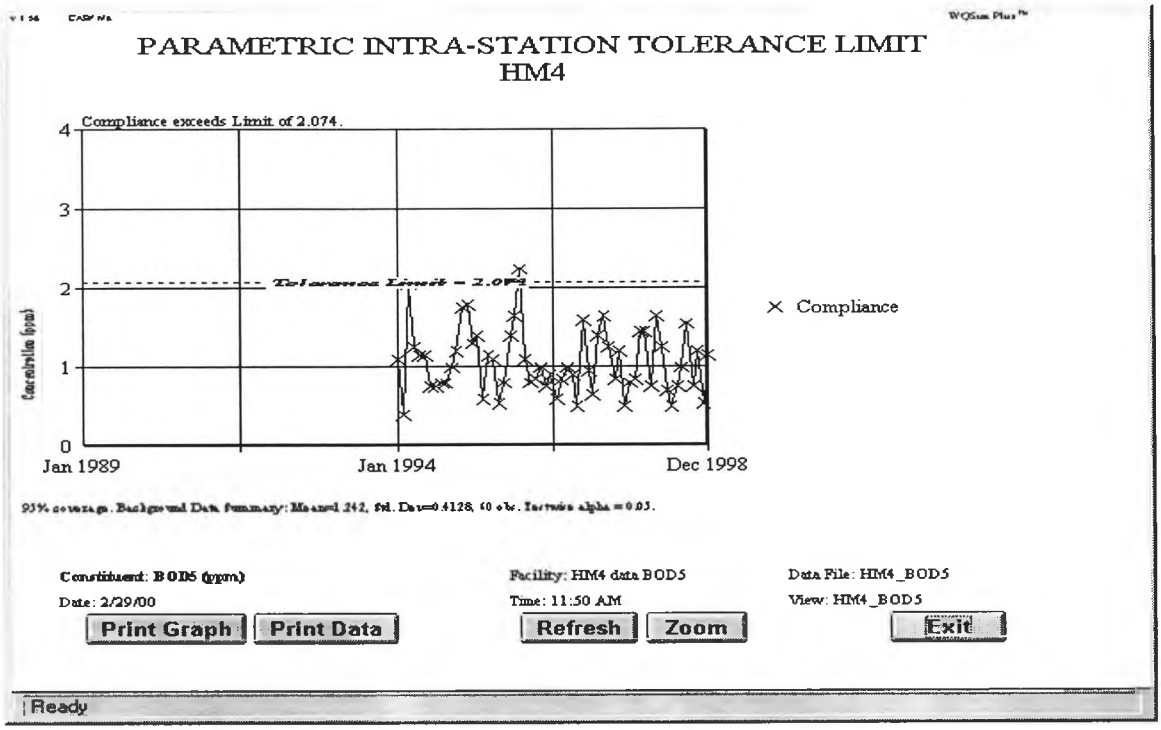


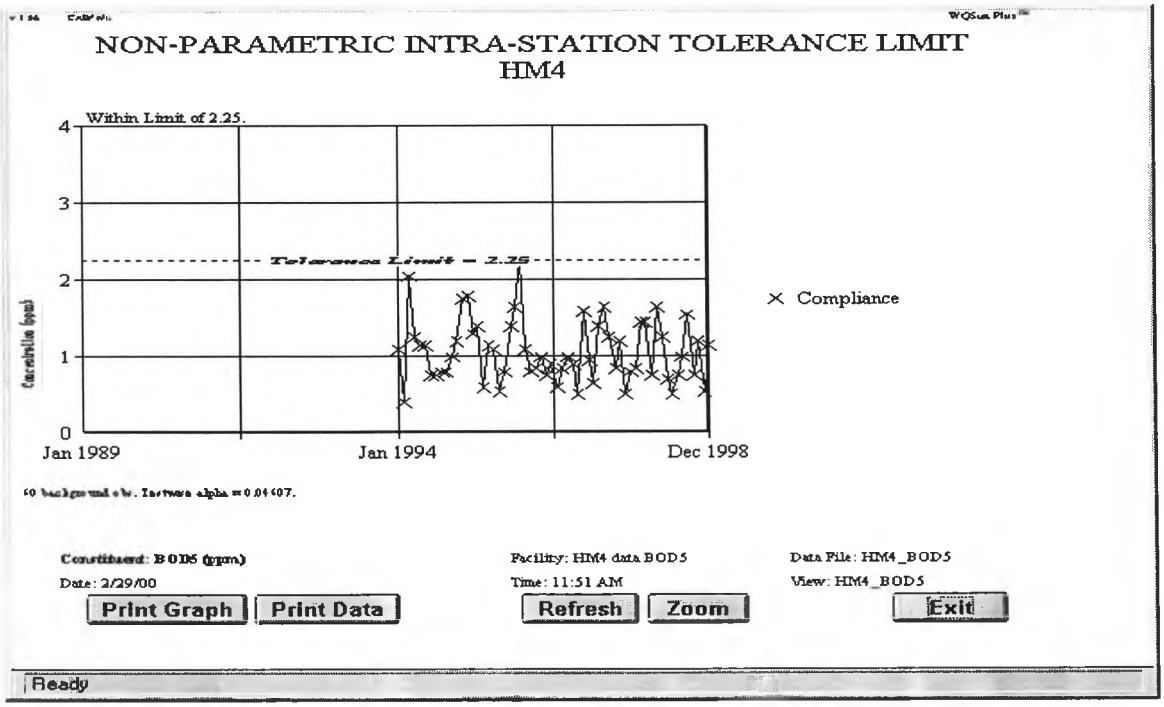
FAC data



Appendix J. Standards Compliance for HM4_BOD5

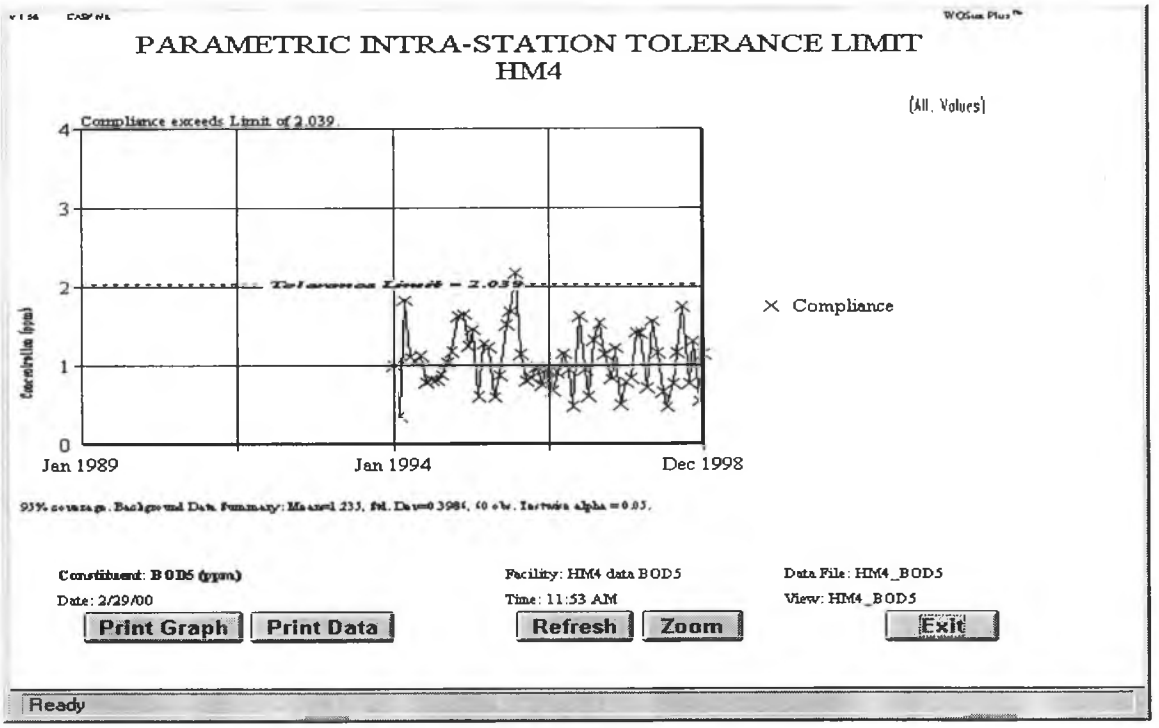
Raw data

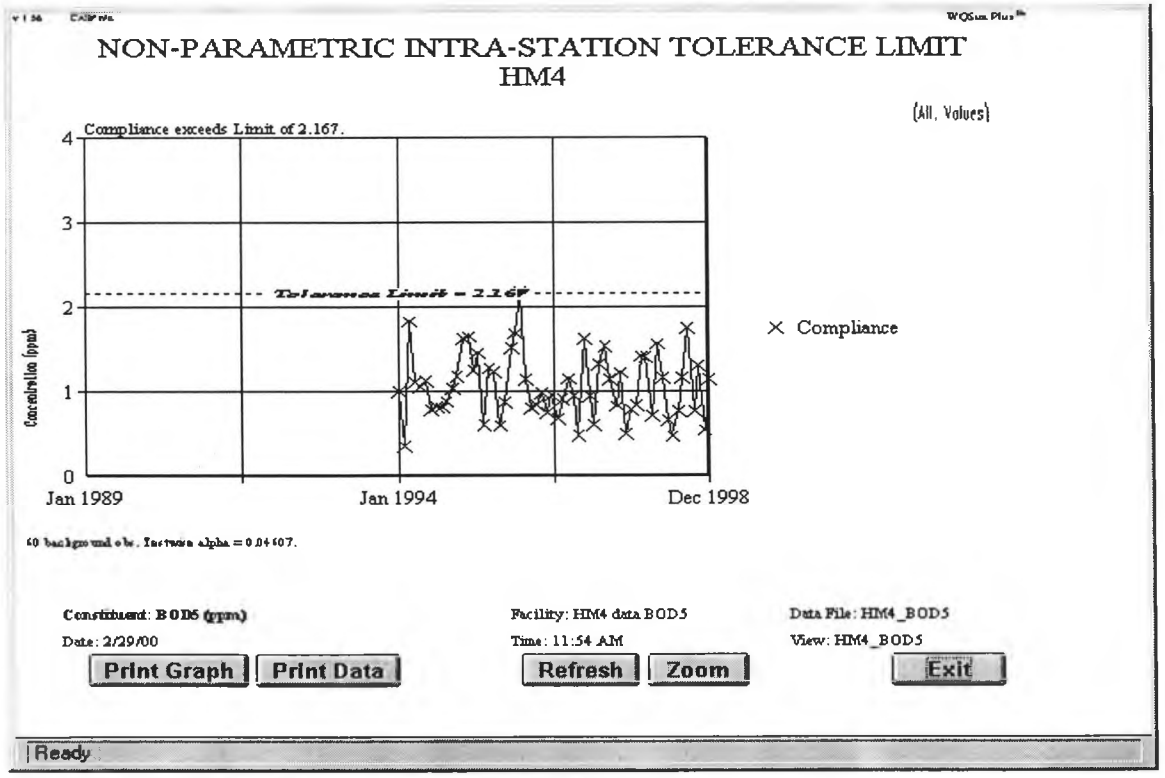




Appendix J. Standards Compliance for HM4_BOD5

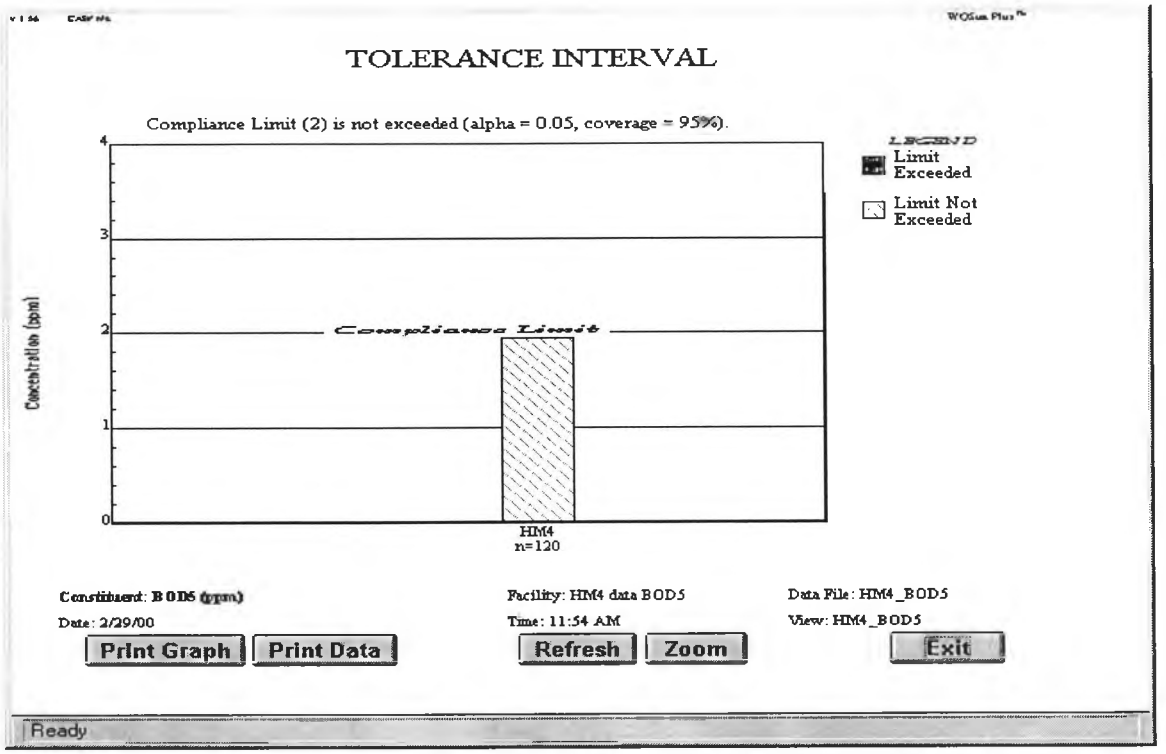
FAC data

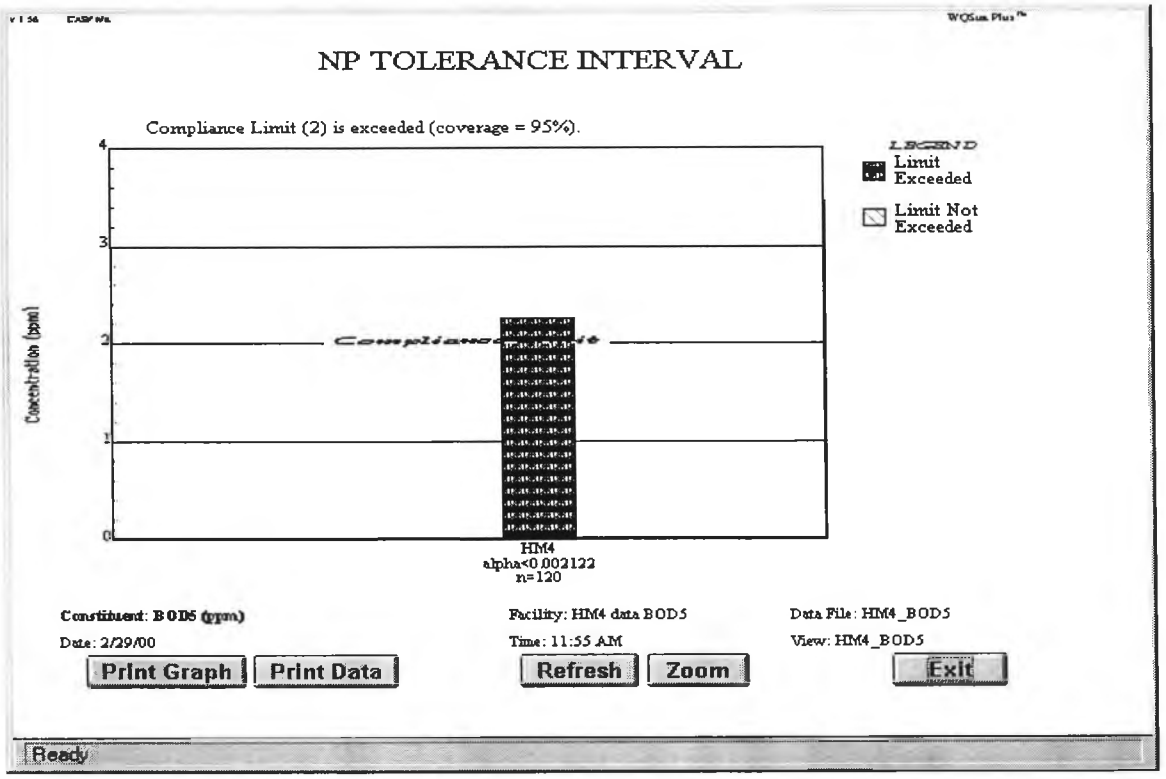




Appendix J. Standards Compliance for HM4_BOD5

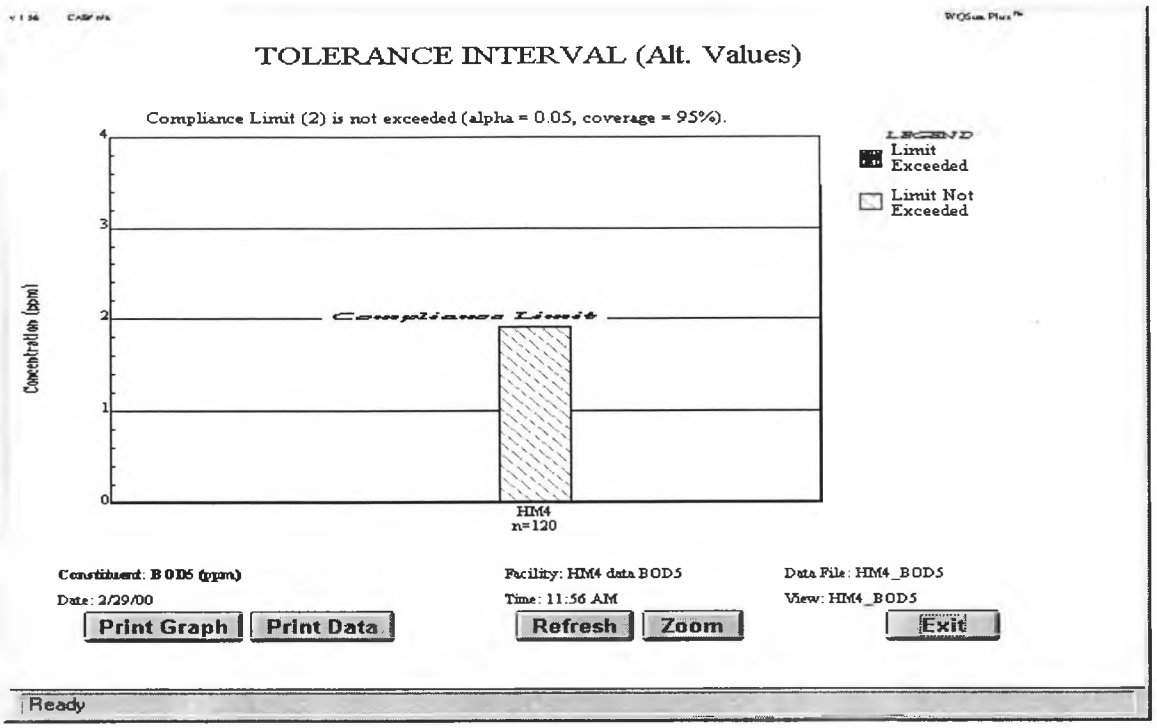
Raw data

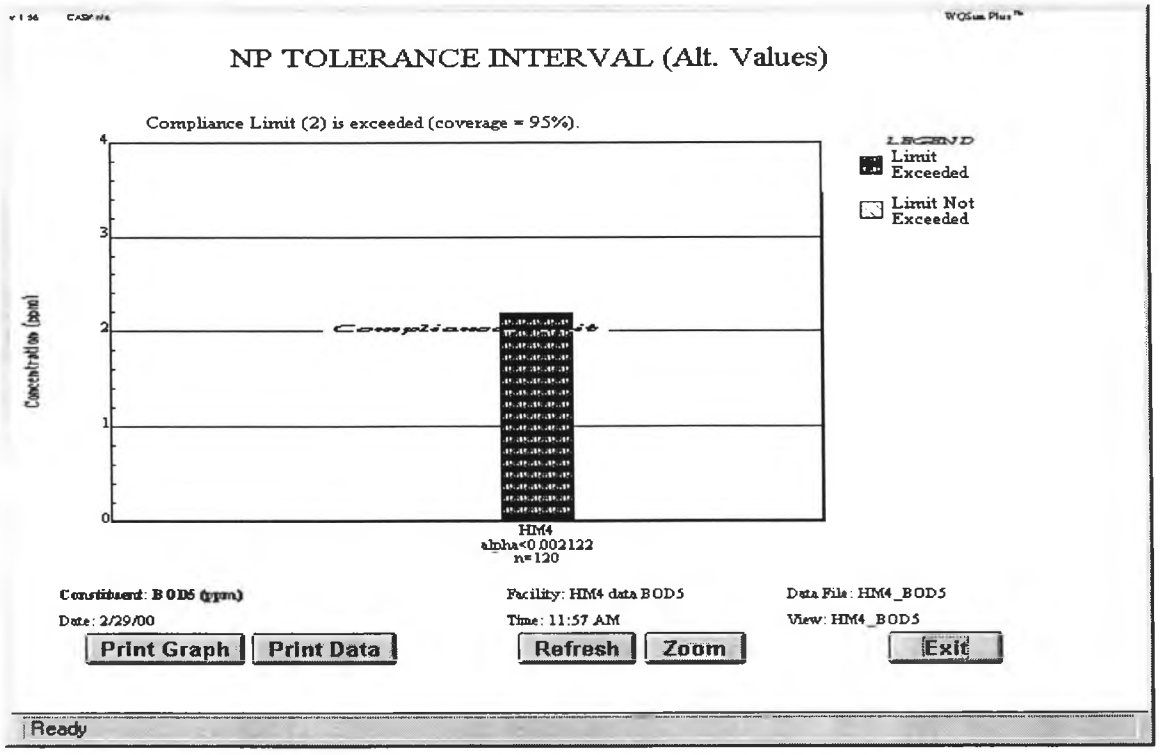




Appendix J. Standards Compliance for HM4_BOD5

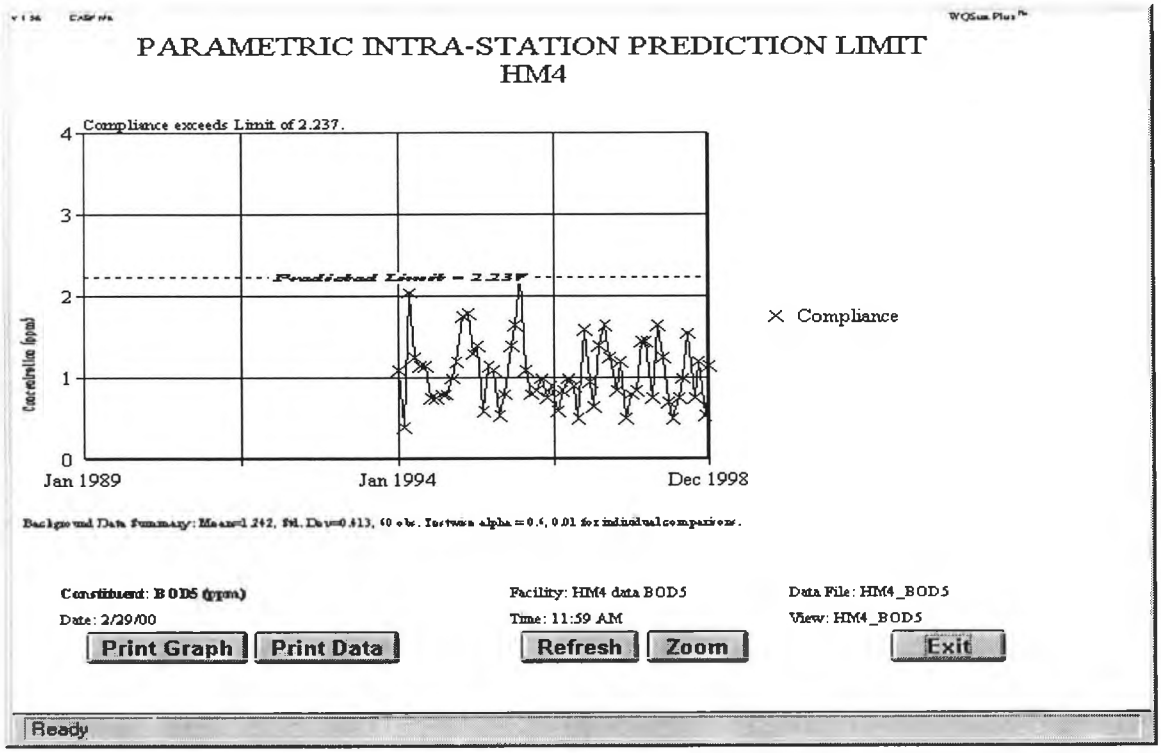
FAC data

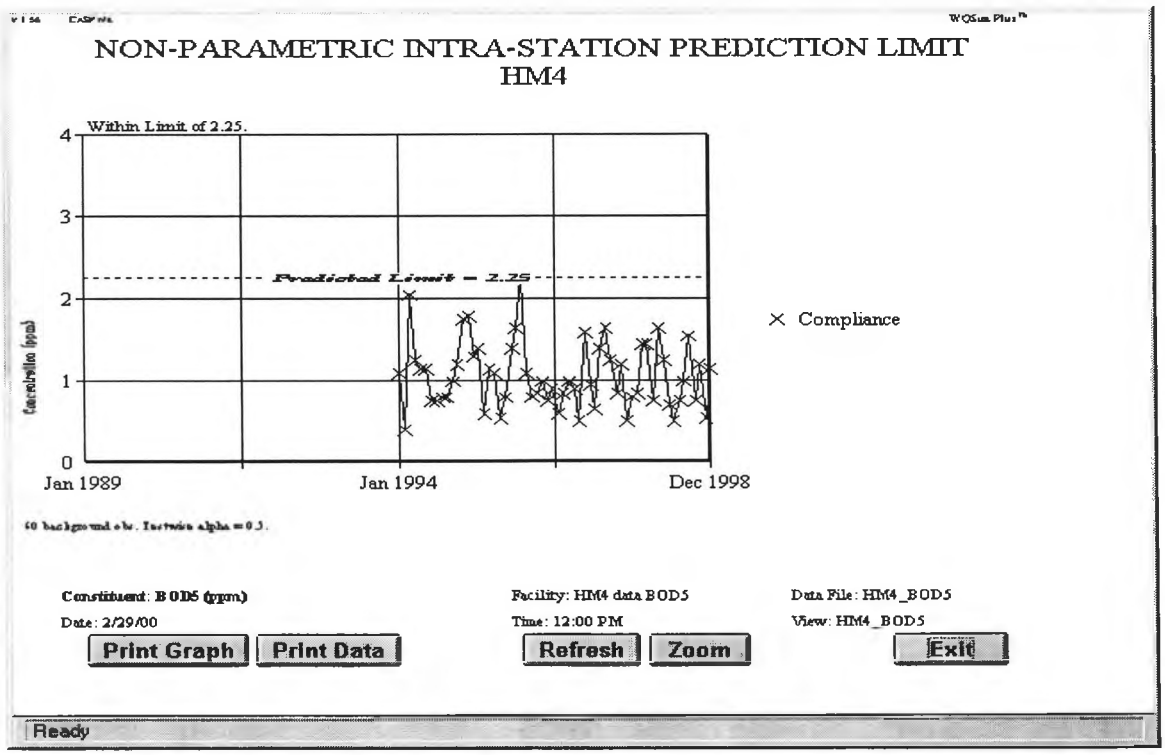




Appendix J. Standards Compliance for HM4_BOD5

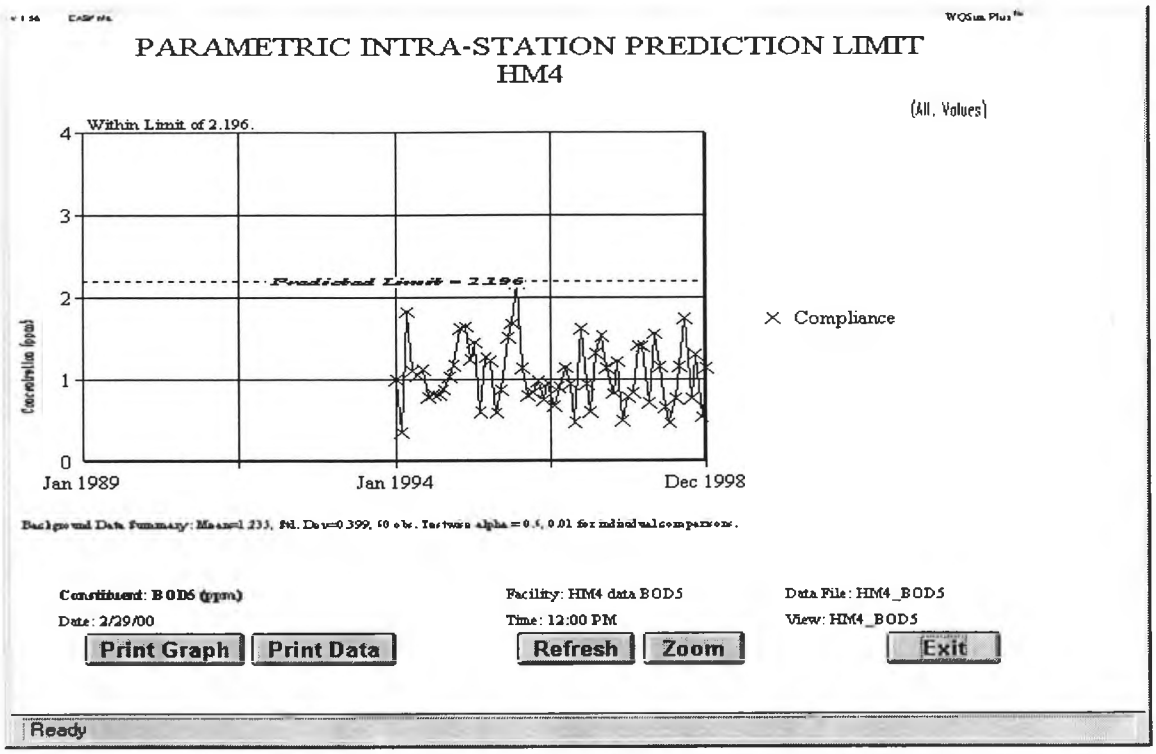
Raw data

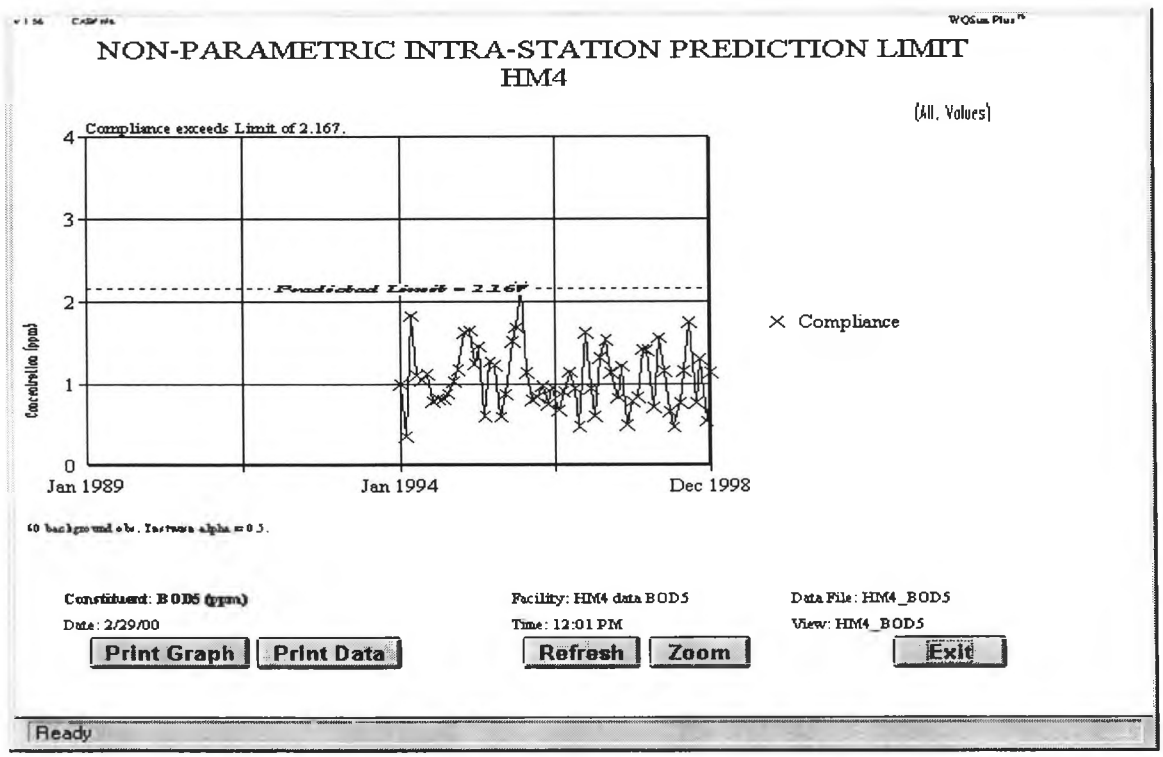




Appendix J. Standards Compliance for HM4_BOD5

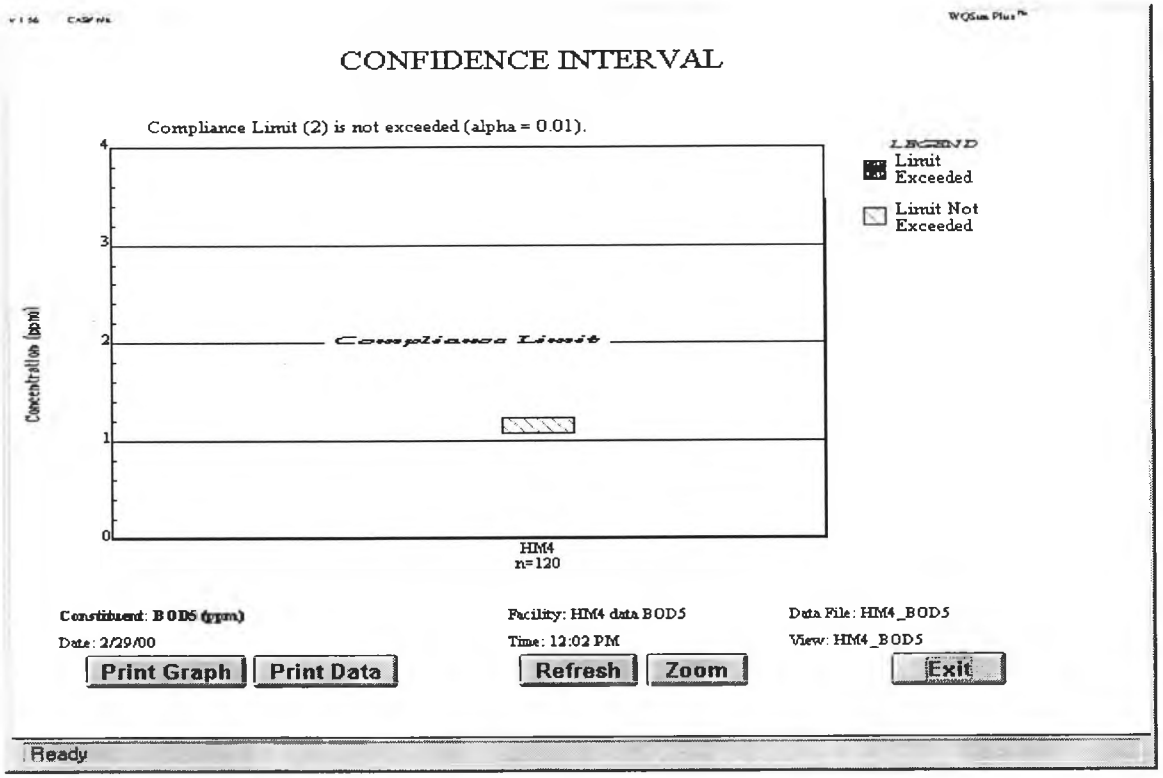
FAC data

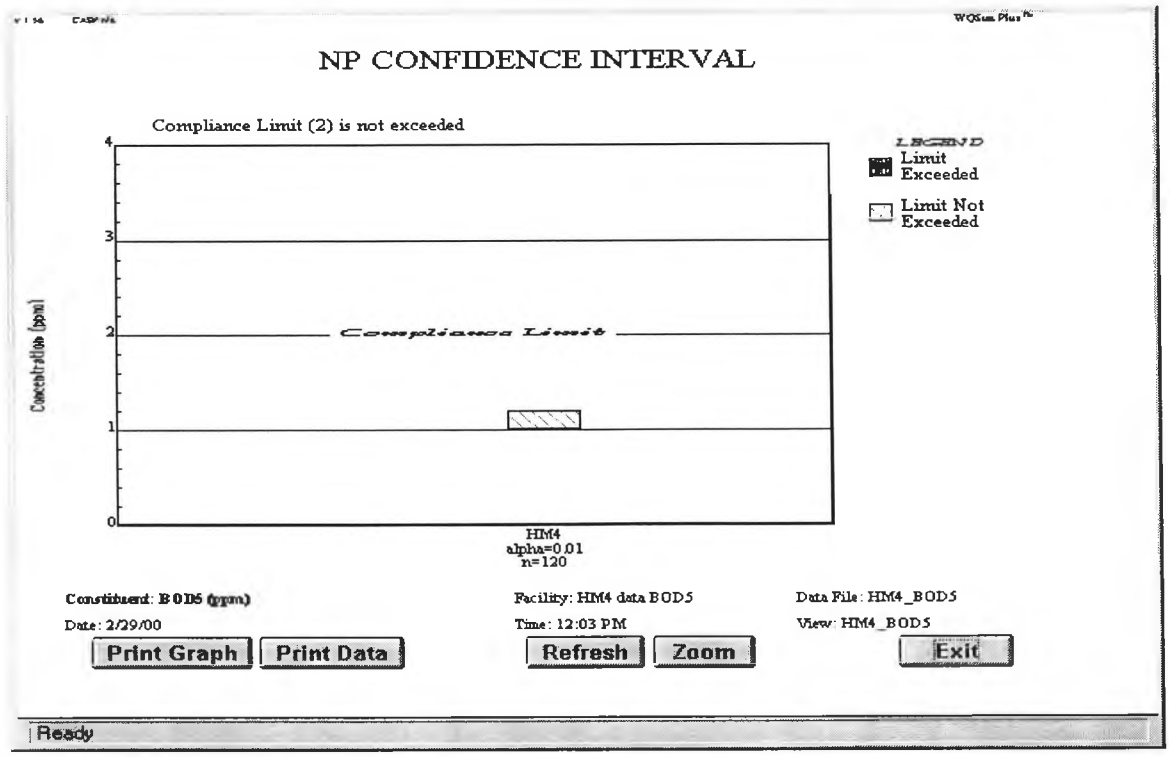




Appendix J. Standards Compliance for HM4_BOD5

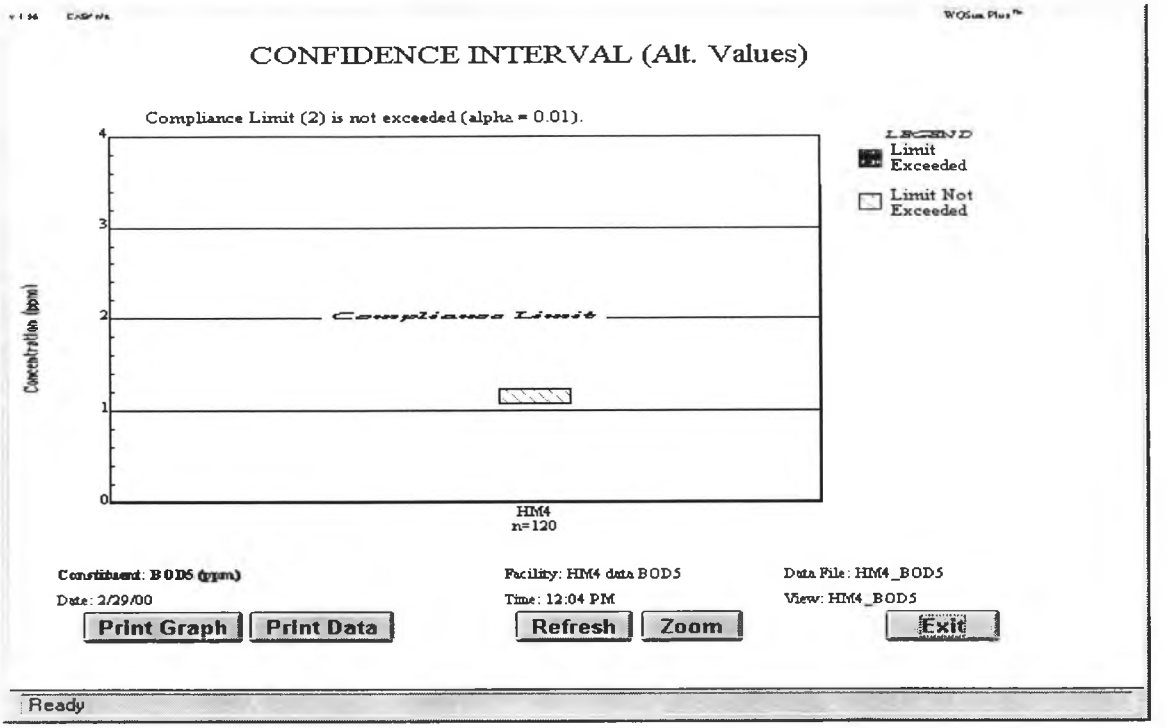
Raw data

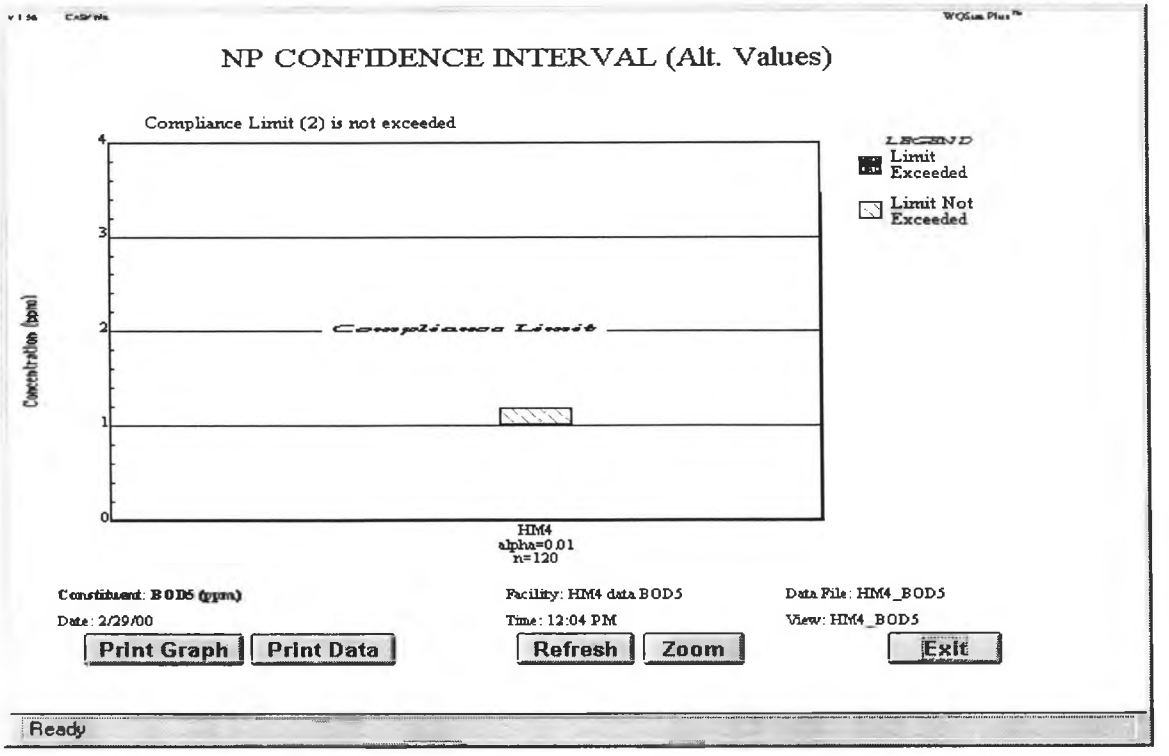




Appendix J. Standards Compliance for HM4_BOD5

FAC data

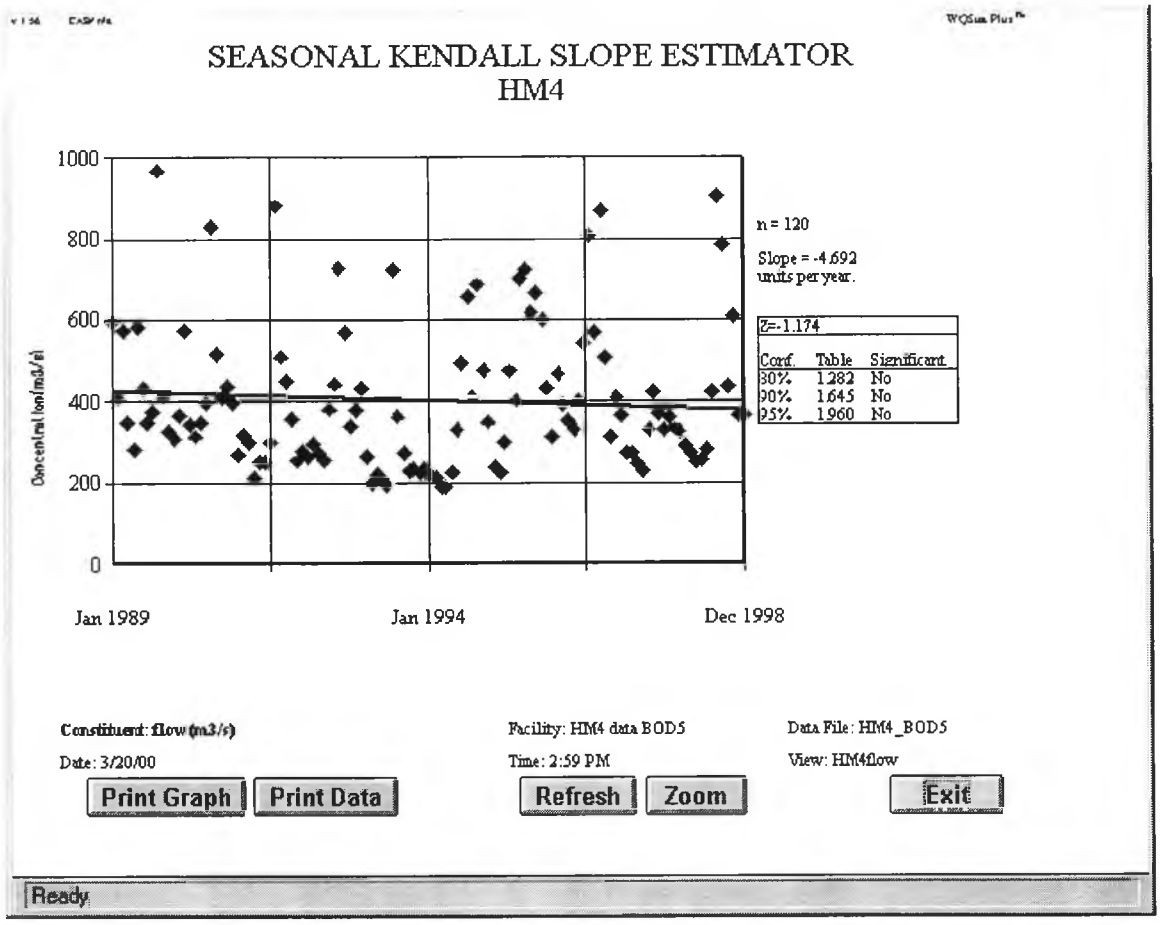




Appendix K.
Trend Results for Flow Data using Seasonal Kendall Test for Trend

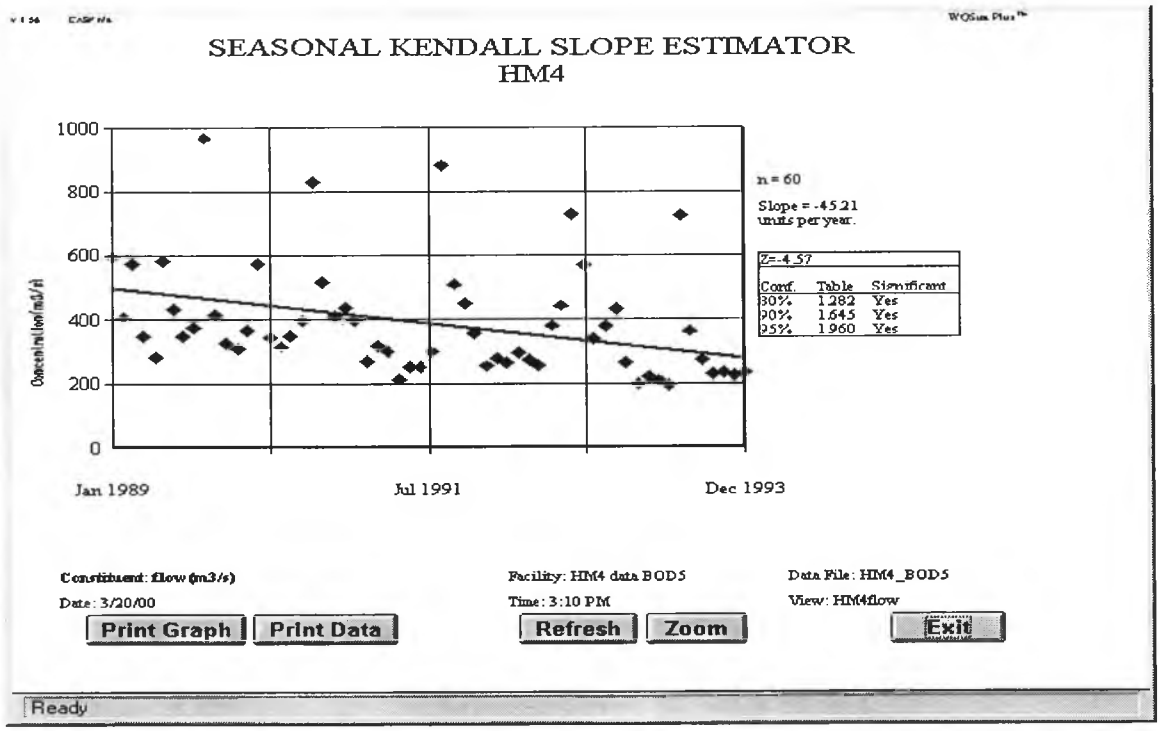
Appendix K1. Trend Results on Flow Data for HM4

10-yr data

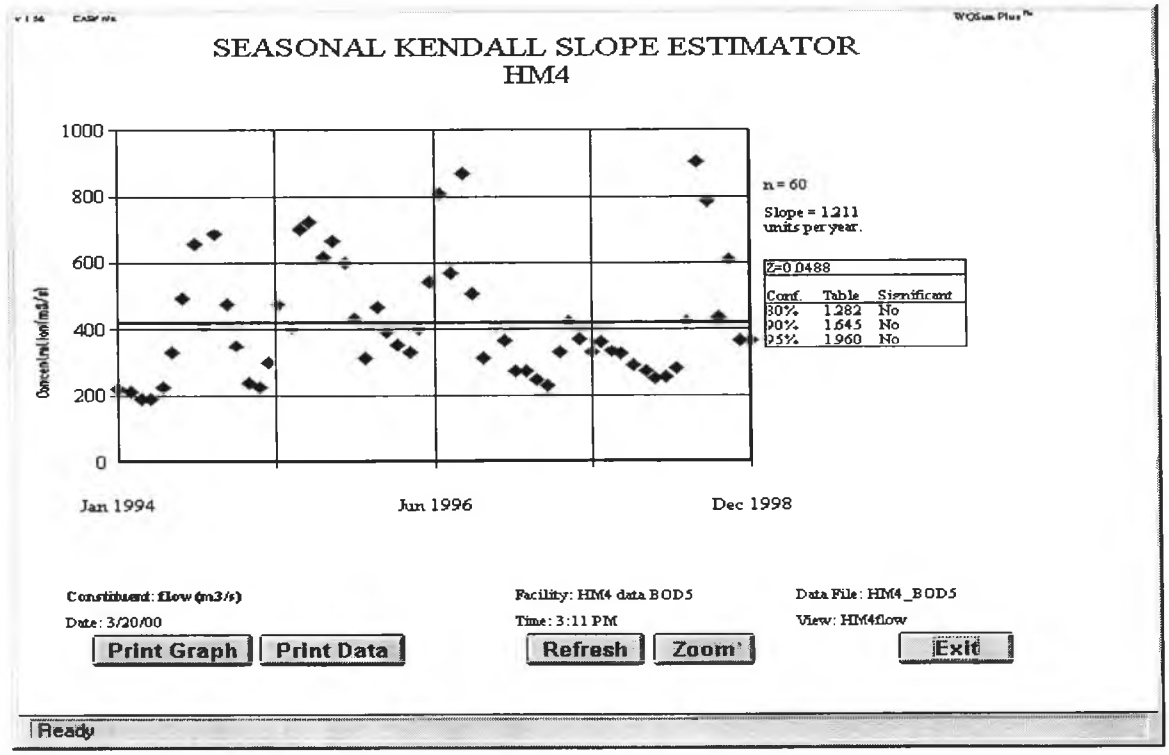


Appendix K1. Trend Results on Flow Data for HM4

1st 5-yr data

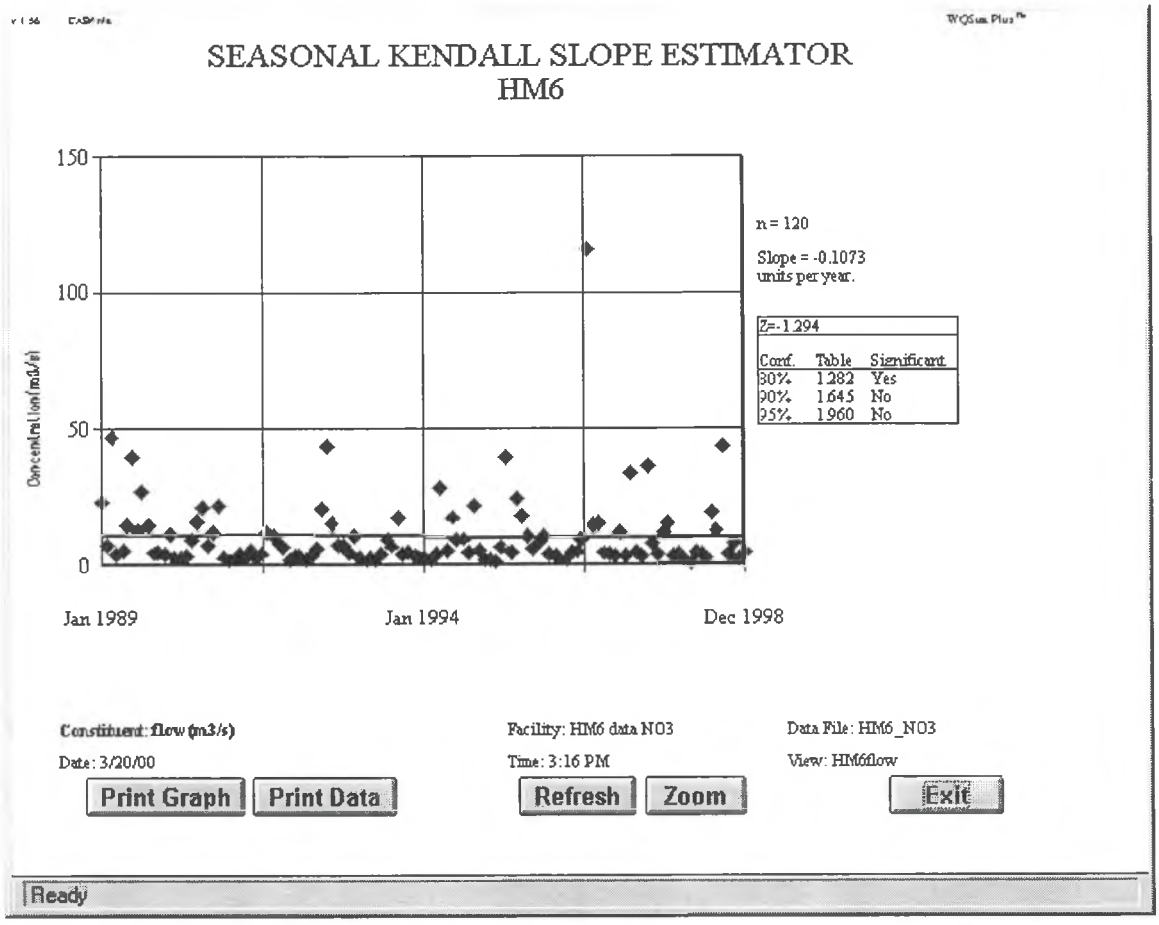


2nd 5-yr data



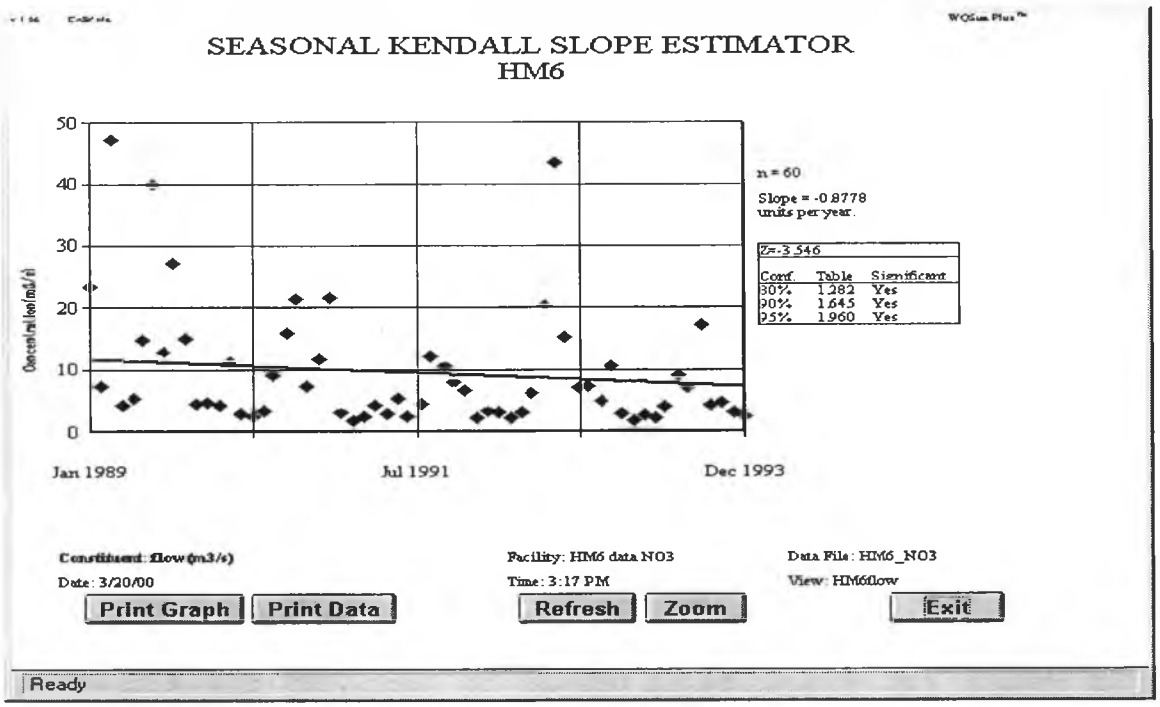
Appendix K2. Trend Results on Flow Data for HM6

10-yr data

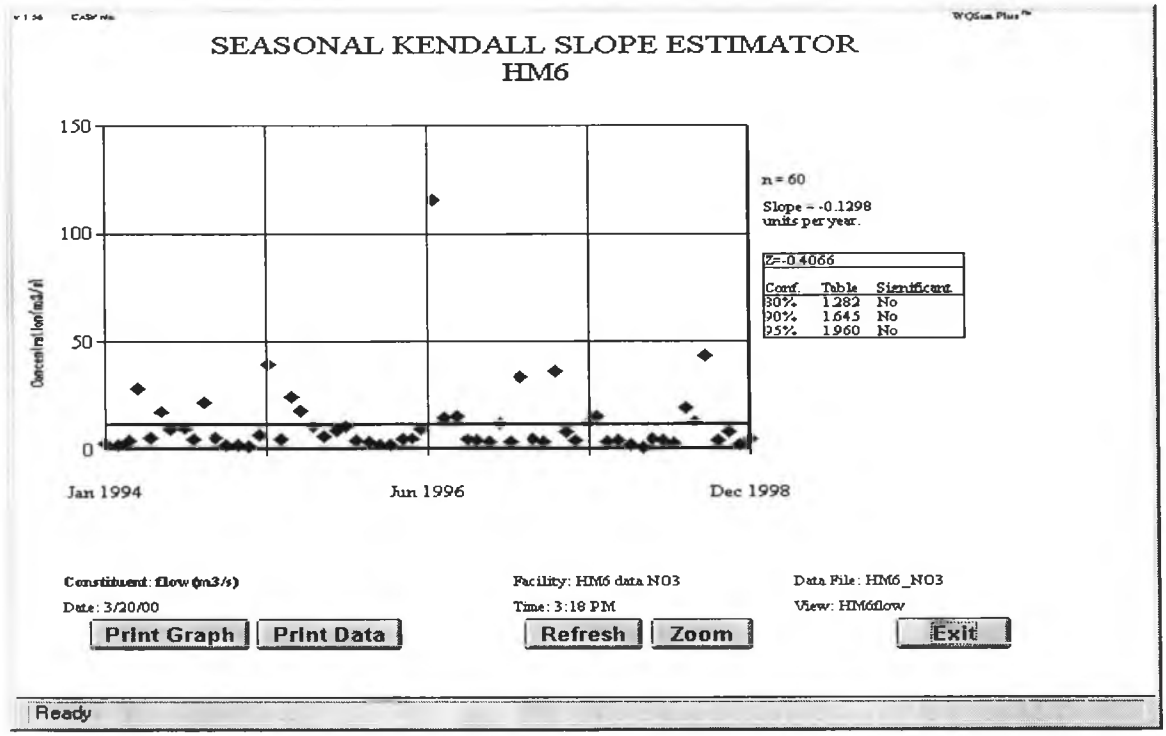


Appendix K2. Trend Results on Flow for HM6

1st 5-yr data

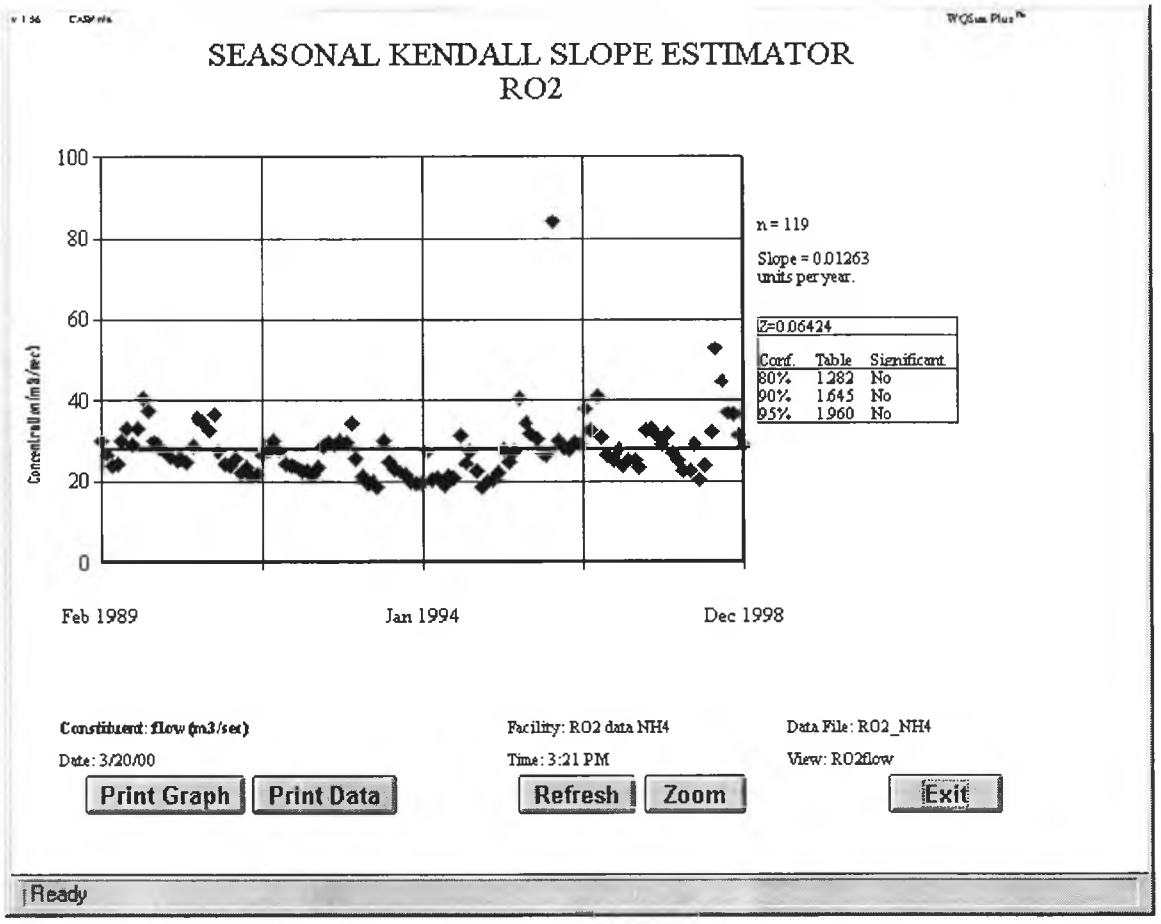


2nd 5-yr data



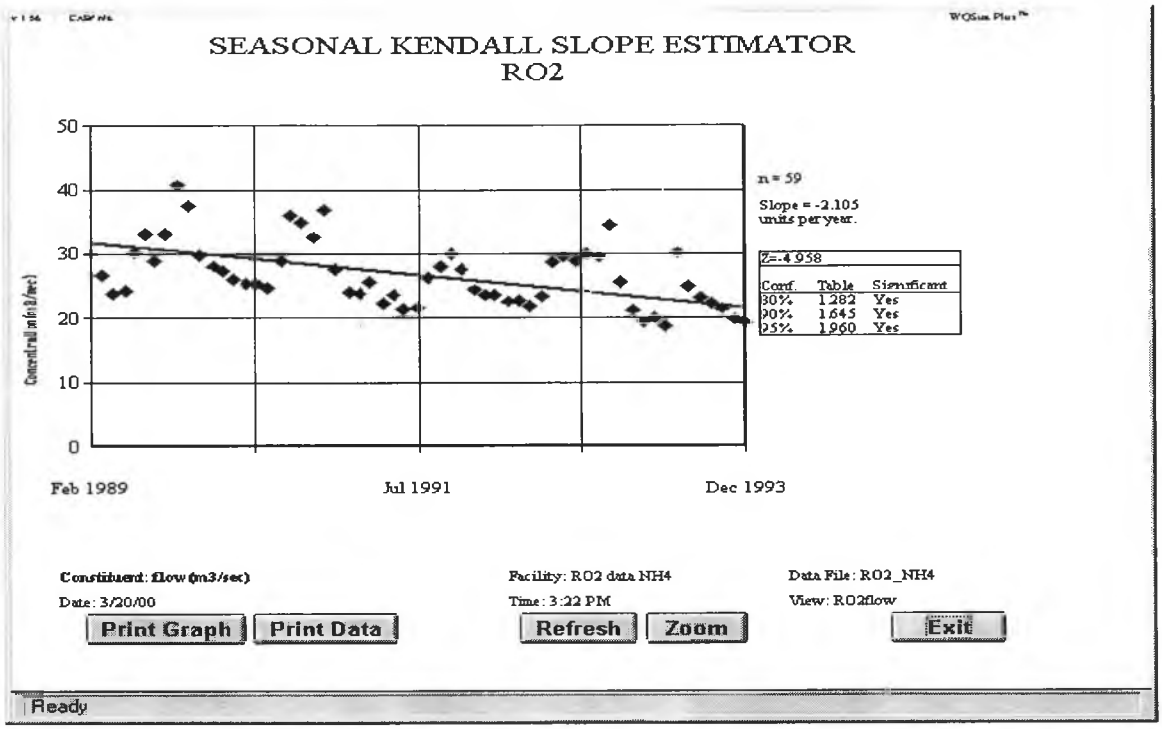
Appendix K3. Trend Results on Flow Data for RO2

10-yr data



Appendix K3. Trend Results on Flow for RO2

1st 5-yr data



2nd 5-yr data