Consolidated Assessment and Listing Methodology

Toward a Compendium of Best Practices

First Edition

July 2002

Prepared By:
U.S. Environmental Protection Agency
Office of Wetlands, Oceans, and Watersheds

With Assistance From:
EPA Regional Offices
Office of Science and Technology
Office of Research and Development
Office of Ground Water and Drinking Water
Office of Wastewater Management
Office of Environmental Information
Office of General Counsel

Disclaimer

This document provides guidance to EPA and states, territories and authorized tribes regarding water quality monitoring and assessment programs. This document does not create any legally binding requirements, but rather suggests approaches that may be used as appropriate. This document does not substitute for EPA's statutes and regulations, and interested parties are free to raise questions and objections about the appropriateness of the application of the examples presented in this guidance to a particular situation. EPA may change this guidance in the future.

Contents

1. Introduction

Contents

1. Introduction

This document outlines an iterative process, a series of steps, for improving states', territories', and authorized tribes' monitoring and assessment programs. The first step is documentation of the decision making process for attainment of water quality standards (WQS) and making that process available to the public. The second step is ensuring that monitoring designs provide data to support management decisions, including the WQS attainment status of waters. The third step is updating decision making methodologies as more high-quality data become available.

This document will provide information to states and other jurisdictions responsible for collecting data and information on water quality that are used for

- Determining the status of attainment for waters within their jurisdiction

- Identifying waters that are impaired and need to be included in Category 5 of the *Integrated Water Quality Monitoring and Assessment Report* (i.e., the Clean Water Act Section 303(d) list). Hereinafter Integrated Water Quality Monitoring and Assessment Report is referred to as the Integrated Report.

1.1 What Is the Objective of This Document?

The immediate objective of this document is to provide a framework for states, territories, interstate commissions, and authorized tribes to document the decision making processes used to assess WQS attainment. This framework includes the organizational structure for documenting the state's assessment and listing methodology and also provides information on the content of these methodologies. For example, it describes each of the types of data that support water quality decision making and how they are used to support different water quality determinations.

In the short term, this framework will promote better documented water quality assessments and greater transparency in decision making about WQS attainment, and will foster greater participation among organizations involved in water quality monitoring and assessment. Over the long term, these efforts should result in more comprehensive, more efficient, and more effective water quality monitoring programs. Clearly, this is an ongoing process, involving continual fine-tuning and improvement of not just the states' water quality assessment methodologies and monitoring programs, but also of the framework and information on methodologies, to keep pace with advances in water quality assessment techniques and increasing technical expertise.

CALM does not attempt to reproduce the volumes of existing technical guidance on water quality monitoring. Instead, it builds upon the previous efforts of the State/EPA 305(b) consistency workgroup and presents a framework for integrating old and new guidance documents into a consolidated monitoring, assessment, and listing methodology. Wherever possible, this document includes citations (and links to web pages) to additional references and resources on data quality, data interpretation, monitoring design, and other technical issues related to water quality assessments and listing decisions. This approach encourages the functional integration of monitoring, data documentation and sharing, and data analysis and

interpretation, among state programs and other partners. It is also designed to encourage consistency of program implementation among EPA regions throughout the country.

1.2 Organization and Format of the Document

This document is formatted as a series of questions that states, territories, interstate commissions, and authorized tribes need to answer to document their current methodology. For each of these questions, the document provides some context about why the question is relevant and some examples of appropriate ways to answer it. The examples are drawn primarily from existing guidance and state programs or proposals. The questions may already be addressed through existing state, territory, tribal monitoring strategy documents, quality assurance project plans, and/or WQS implementation procedures. To the extent these other documents describe the assessment and listing methodology, the states' work is essentially done and can merely be cross-referenced or compiled into a single assessment and listing methodology.

The remainder of this document is organized into three parts. Part A deals with WQS attainment decisions and identification of impaired waters and is organized according to the types of data that may be used to support these decisions. Within each of these chapters, the document sets forth questions for states about how they define data quality requirements and how they utilize and interpret data to make decisions about whether a water is impaired or WQS have been attained.

Part B deals with designing a comprehensive monitoring program to assess the extent to which waters are attaining WQS and to identify the waters that are impaired. This part addresses the overall design of water quality monitoring programs, including documenting monitoring goals and data quality objectives for the type, amount, and scale of data needed. One chapter explores options for extending monitoring programs over time to cover all water resource types, including lakes, rivers, wetlands, estuaries, and coastal waters. It presents information on using probability-based sampling design to generate statewide characterizations of the extent of waters attaining WQS or impaired. Another chapter describes a targeted or followup stage of sampling designs for making attainment/impairment decisions about specific drainage areas, waterbodies, or segments.

Part C describes approaches for reporting on WQS attainment for the full inventory of waters in the Integrated Report. This part addresses the documentation used to communicate the findings and the basis of attainment/impairment decisions. It provides different options for presenting findings at different scales relevant to the sampling design. For example, the Integrated Report may contain one section that presents the overall extent of water quality conditions based on statewide probability designs, followed by a series of watershed or basin-level sections that present the results of finer scale monitoring to identify impaired waters based on site-specific information.

The question-and-answer format of this document provides a framework for a consolidated assessment and listing methodology, as well as information, including examples, about ways to

respond to the questions. The examples given are not exclusive, so that flexibility is allowed to reflect legitimate variations among states, territories, and authorized tribes in the WQS adopted and assessment methodologies employed. Not all states, territories, and authorized tribes currently have programs that reflect the information and examples described in the document. EPA regional office staff will encourage states, territories, and authorized tribes to define the improvements needed in their programs, and to develop an implementation plan and timeline for making these changes.

1.3 References

U.S. EPA. 2001. 2002 Integrated Water Quality Monitoring and Assessment Report Guidance Memorandum. Robert H. Wayland III, Director, Office of Wetlands, Oceans and Watersheds. November 19, 2001.

Part A — Water Quality Standards Attainment Decisions

Contents

3. Overview of Process To Assess WQS Attainment Status and Identify Impaired Waters

Contents

3. Overview of Process To Assess WQS Attainment Status and Identify Impaired Waters

3.1 Introduction

Most states, territories, and authorized tribes organize their water quality standards (WQS) according to the designated beneficial uses assigned to waters. Recall that the WQS consists of three elements, the designated use, the narrative and numeric criteria adopted to protect the use, and antidegradation policies. Once these WQS are adopted, the state ensures they are met. This includes monitoring to assess attainment status, reporting on attainment and identifying impaired waters, and implementing appropriate measures to ensure WQS are met.

For each WQS, the state, territory, or authorized tribe should describe how it assesses attainment with the standard. The description may be included in the approved WQS or in other implementing regulations or policies and procedures such as the state, territory, or authorized tribe's continuous planning process or consolidated assessment and listing methodology. This includes defining the water quality indicators it measures and the procedures for analyzing and interpreting data in order to decide whether standards are met or water quality is impaired. This should include collection and analysis of multiple types of data providing information relevant to assessing attainment with approved WQS. This information not only is used for reporting attainment status in the Integrated Report but also supports development of appropriate controls that address the full range of water quality problems.

This chapter is organized according to general categories of designated use–based WQS: aquatic life, recreation, public water supply, and fish and shellfish consumption. Each section briefly describes the types of data frequently used in WQS attainment decisions and how these data are interpreted. It also presents examples of how states, territories, and authorized tribes work through situations in which different data types do not indicate the same attainment decision.

3.1.1 Elements of State Water Quality Standards

The objective of the Clean Water Act (CWA) is to "restore and maintain the physical, chemical, and biological integrity of the Nation's waters." To achieve this objective, section 303(c)(2) calls for states, territories, and authorized tribes to adopt WQS including designated uses, narrative and numeric criteria to protect those uses, and antidegradation policies to prevent deterioration of high-quality waters. Under section 106(e), states, territories, and authorized tribes also implement monitoring programs that allow them to report on attainment of WQS and to identify and prioritize waters not attaining standards.

Section 101(a)(2) of the CWA establishes as a national goal "water quality which provides for the protection and propagation of fish, shellfish, and wildlife, and recreation in and on the water, wherever attainable." Section 303(c)(2)(A) of the CWA requires WQS to protect the public health and welfare, enhance the quality of water, and serve the purposes of the Act. EPA's regulations at 40 CFR 131 interpret and implement sections 101(a) and 303(c)(2)(A) of the CWA by requiring that State WQS provide at a minimum for the section 101(a) "fishable/swimmable" uses unless those uses have been shown to be unattainable. In designating

waters, states, territories, and authorized tribes consider the use and value of water for public water supplies; protection and propagation of fish, shellfish and wildlife; recreation in and on the water; consumption of fish and shellfish by humans; and agricultural, industrial, and other purposes including navigation. In no case may waste transport or assimilation be adopted as a designated use for any waters of the United States. Table 3-1 is an example of the many designated uses that may be adopted by a state.

Table 3-1. California's applicable designated uses

Agricultural supply	Marine habitat*
Aquaculture	Municipal and domestic navigation
Cold freshwater habitat*	Noncontact recreation
Commercial and sport fishing*	Preservation of biological habitats of special
Estuarine habitat*	significance*
Fish spawning*	Rare and endangered species*
Fish migration*	Saline water habitat*
Flood control	Shellfish harvesting*
Freshwater replenishment	Warm freshwater habitat
Groundwater recharge	Water quality enhancement
Hydroelectric power generation	Water contact recreation
Industrial service supply	Wildlife habitat
Industrial process supply	

* Aquatic life–related uses.

Although some states, territories and authorized tribes have detailed categories and subcategories of designated uses that apply to specific waters or classes of waters, many have adopted general categories of use that apply broadly to all waters. A recent report by the National Research Council recommended that states, territories, and authorized tribes move beyond general categories of "fishable" and "swimmable" and adopt refined or detailed uses that better describe the expectations for the water (NRC 2001). For example, a state, territory, or authorized tribe may want to distinguish between primary contact recreation and secondary contact recreation. Similarly, the aquatic life use should describe the attributes of aquatic communities expected for the water.

States, territories, and authorized tribes adopt numeric and narrative water quality criteria to protect designated uses. Numeric water quality criteria are adopted based on EPA's 304(a) criteria guidance, 304(a) criteria modified to reflect site-specific conditions, or other scientifically defensible methods. Narrative criteria are adopted to supplement numeric criteria or if numerical criteria cannot be determined. Narrative criteria are descriptions of the conditions necessary for a waterbody to attain its designated use, whereas numeric criteria are values expressed as chemical concentrations, toxicity units, aquatic community index levels, or other numbers deemed necessary to protect designated uses. A "translator" identifies a process, methodology, or guidance that States or Tribes will use to quantitatively interpret narrative criteria statements. Translators may consist of biological assessment methods (e.g., field measures of the biological community), biological monitoring methods (e.g., laboratory toxicity tests), models or formulae that use input of site-specific information/data, or other scientifically defensible methods. Translators are particularly useful for addressing water quality conditions that require a greater degree of sophistication to assess than can be typically expressed by

numerical criteria that apply broadly to all waters with a given use designation. Criteria must be based on sound scientific rationale and should contain sufficient parameters or constituents to protect the designated use. The National Research Council report also emphasized selection of criteria that are accurate indicators of the designated use.

Where a state, territory or authorized tribe adopts narrative criteria for toxic pollutants to protect designated uses, it must provide information identifying the method by which it intends to regulate point source discharges of toxic pollutants on water quality limited segments based on such narrative criteria. Such information may be included as part of the standards or may be included in documents generated by the state, territory or authorized tribe in response to the Water Quality Planning and Management Regulations (40 CFR part 35). Where a state, territory, or authorized tribe adopts narrative criteria for non-toxic pollutants to protect designated uses, it should provide information identifying the method by which it intends to regulate point sources discharges on water quality limited segments based on such narrative criteria in the state, territory, or authorized tribe's WQS or alternatively in other implementing regulations or policies and procedures documents such as the continuous planning process of consolidated assessment and listing methodology.

States, territories, and authorized tribes also adopt an antidegradation policy specifying the framework to be used in making decisions regarding changes in water quality. The intent of an antidegradation policy is to ensure that in all cases, at a minimum: (1) water quality necessary to support existing uses is maintained; (2) where water quality is better than the minimum level necessary to support protection and propagation of fish, shellfish, and wildlife, and recreation in and on the water ("fishable/swimmable"), that water quality is also maintained and protected unless, through a public process, some lowering of water quality is deemed necessary to allow important economic or social development to occur; and (3) where waterbodies are of exceptional recreational or ecological significance, water quality is maintained and protected.

3.1.2 Monitoring To Assess Attainment With WQS and Identify Impaired Waters

Monitoring to determine attainment of applicable WQS should include a multi-indicator approach that may include biological, toxicological, physical, and chemical indicators of the WQS and its components. Each type of data provides unique insights into the integrity and health of an aquatic system, as well as the ability of the public to safely recreate in such waters. These indicators are frequently organized on a continuum from loadings entering the system, to stressors present in the system, to response of the system (see Figure 3-1).

Each type of data offers different strengths and limitations. For example, biological assessments measure the response of aquatic life to the cumulative effects of past or current impacts from multiple physical and chemical stressors. However, these assessments may be limited in their ability to predict future impacts, or identify new stresses that have not begun to be reflected in the biological community. Chemical-specific assessments evaluate and predict impacts from single pollutants, but do not capture the combined interactions of pollutants or their cumulative impacts over time. Assessment of the physical, chemical, and biological integrity of the nation's waters should be based on an appropriate combination of indicators selected to characterize WQS

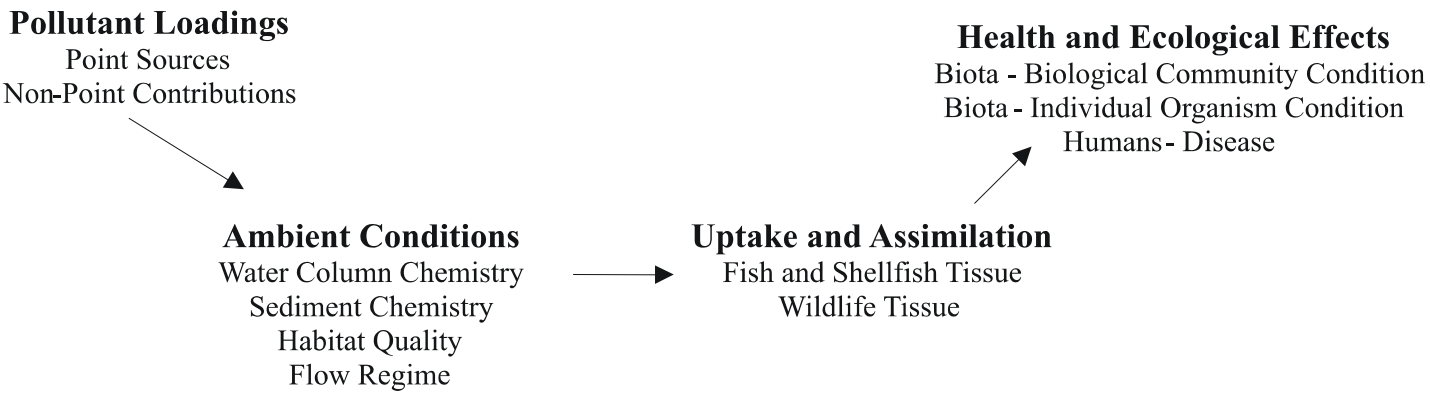


Figure 3-1. Continuum of water quality indicators.

attainment status, including physical, chemical, and toxic characteristics of water and sediment; chemical accumulations in fish tissue; a biological assessment of the aquatic community; and physical condition of habitats. Chapter 10 provides more information on selection of indicators.

State, territory, or tribal WQS (uses, criteria, and the antidegradation policy) adopted pursuant to section 303(c) of the CWA are the basis for attainment or nonattainment determinations for the purposes of identifying impaired waters pursuant to CWA section 303(d). Under section 303(d)(1) of the CWA, states, territories, and authorized tribes must identify waterbodies for which technology-based controls required by the Act are not sufficient to implement applicable WQS, and prioritize such waterbodies for TMDL establishment. For purposes of determining whether a waterbody is impaired and should be included on section 303(d) lists, states, territories, and authorized tribes are required by EPA regulations to consider all existing and readily available data and information. For example, if a state shares a waterbody with another state, it must consider existing and readily available data from the state that shares the waterbody. This may include physical, chemical, and biological data, including data on pathogens (such as bacteria and phytotoxins), as well as fish and shellfish tissue concentration data, where such data are existing and readily available. The assessment methodology prepared by states, territories, and authorized tribes should describe how it collects or obtains data and information relevant to applicable WQS, how it evaluates the suitability of the data or information for decision making, and how it analyzes and interprets data to make attainment or nonattainment decisions.

3.2 Aquatic Life–Based Water Quality Standards

Section 101(a)(2) of the CWA establishes as a national goal "water quality which provides for the protection and propagation of fish, shellfish, and wildlife, and recreation in and on the water, wherever attainable." EPA's WQS regulations require that standards provide for these "fishable/swimmable" uses wherever attainable. Each state, territory and authorized tribe develops and adopts aquatic life–based WQS for waters under its jurisdiction.

The form of aquatic life–based WQS varies from state to state. Some states adopt aquatic life–based WQS that contain broad aquatic life uses and criteria that essentially apply to all waters. Others, like the example in Table 3-1, have a variety of different aquatic life use categories based on the type or function of the water. The narrative and numeric criteria adopted to protect those uses may apply to all of the uses or be tailored to each specific use. Still others have used biological assessments to develop refined or tiered aquatic life–based WQS categories that reflect expectations for characteristics of the aquatic community in each category. With tiered aquatic life uses, a state can set numeric biocriteria that clearly define the upper and lower bounds of biological conditions expected within each aquatic life tier. Similiarly, the state can adopt physical and chemical criteria appropriate for each tier. When approached in this fashion, a state will have aquatic life–based WQS that clearly and precisely define what the management objective is for a given waterbody and the numeric thresholds or criteria above and below which the objective is or is not achieved.

Some states, such as Maine, Ohio, Vermont, Florida, Maryland, Kentucky, and Oregon, have already constructed biological assessment and standards programs for streams and small rivers incorporating tiered aquatic life uses derived from their bioassessment data, and are protecting those uses through numeric or narrative biocriteria. Most other states are developing programs and are at different levels of implementation.

3.2.1 Which Types of Data and Information Does the State Use for Assessing Whether Aquatic Life–Based WQS Are Attained?

State water quality standards are comprised of three distinct elements: (1) designated use, (2) numeric and narrative criteria that protect the use, and (3) antidegradation policy. For each state standard, the state should describe how it assesses attainment with the standard, and each component element. Ideally, this description should be included in the state's water quality standards. Alternatively, it may be defined in other implementing regulations or policies and procedures documents such as the state's continuous planning process or consolidated assessment and listing methodology.

States, territories and authorized tribes should describe the indicators and thresholds that are used to assess attainment status for each WQS. The term "indicators" is used to refer to a wide range of measures of water quality (e.g., physical, chemical, biological, habitat, toxicity, tissue data). "Thresholds" refers to the numeric value or narrative description that distinguishes attainment from impairment. These thresholds may be adopted into the state, territory or authorized tribe's WQS ordefined in other implementing regulations or policy and procedure documents as a translator or implementation procedure for interpreting the WQS.

Following are brief descriptions of the various indicators or types of data a state may use to interpret its aquatic life–based WQS. Subsequent chapters in this document provide detailed descriptions of how these different types of data may be used.

- *Biological data*—Biological data measure actual effects of pollutants on an aquatic community. Biological assessments typically quantify the difference between reference or

expected conditions of aquatic communities and those found at a specific site being evaluated. Reference conditions are the expected biological attributes (e.g., the structure, function, and condition) of the aquatic community in a particular type or class of waterbody. Chapter 5 provides more detail and references to technical documents on the use of biological data to assess WQS attainment/impairment. EPA recommends that states include biological indicators among the core indicators used to assess attainment with aquatic life–based WQS.

- *Habitat data*—Habitat assessments are often conducted in conjunction with biological assessments. A general habitat assessment incorporates physical attributes from microhabitat features such as substrate, velocity, and depth, with waterbody morphology features such as width, sinuosity, flow, or volume, and macrohabitat features such as vegetation and land use. All of these features can be incorporated into an index or summary of overall habitat conditions. Typically, states, territories, and authorized tribes integrate habitat assessments with biological assessments when assessing applicable WQS attainment. These indices are sometimes used independently to determine whether aquatic life uses are being attained. Chapter 8 provides more detail and references to technical documents about development and application of habitat indicators.

- *Toxicity data from water column and sediment*—Ambient water column and sediment toxicity tests are useful for examining the effects of unknown mixtures of chemicals in surface waters. They may also be used to confirm that an observed impairment is not due to chemical or toxicity-related sources. Toxicity thresholds are expressed in terms of "toxic units" that cause toxic effects to aquatic organisms. Toxicity levels are determined by exposing aquatic organisms to water samples. To sensitive aquatic organisms, toxicity testing integrates the biological effects of most chemical stressors present, potentially giving a more accurate estimate of the actual water or sediment quality compared with chemical concentration measurements. Even unknown toxicants are addressed during testing.

  States and tribes may have water or sediment toxicity criteria in numeric form (toxic units) or narrative form ("free from"). Whole effluent toxicity (WET) testing is commonly performed at point-source discharges and can be used to trigger monitoring for toxicity. Chapter 6 provides more detail and references to technical documents about the use of toxicity testing as an indicator of WQS attainment.

- *Chemical and physical data*—Chemical and physical data address toxicants (e.g., priority pollutants and nonpriority pollutants) and physical characteristics (e.g., dissolved oxygen, suspended solids, pH, and temperature) in water and sediments. Chemical and physical data provide direct information about whether specific pollutants are present in amounts that are causing or likely to cause adverse impacts to aquatic organisms.

  EPA has published water quality criteria for the protection of aquatic life for 31 pollutants, under the authority of section 304(a) of the CWA. States, territories, and authorized tribes use these water quality criteria as guidance in adopting water quality criteria into their

WQS. Chapter 4 provides more information on the use of chemical and physical data for determining WQS attainment. As described in Chapter 11, EPA recommends the use of physical and chemical indicators as core and supplemental indicators of aquatic life–based WQS.

An important element of a state's consolidated assessment and listing methodology is a description of how it assesses attainment with its WQS. In the most comprehensive circumstance, the state may measure indicators of the use and all applicable numeric and narrative criteria in addition to ensuring that the antidegradation policy is met. A state following this approach would identify a water as attaining a particular WQS only when the state has demonstrated that all of these indicators are in attainment.

States are often more selective in the water quality indicators used to assess attainment with water quality standards. States may describe a subset or hierarchy of indicators that serve to characterize whether a WQS (and its components) are attained. Under this approach, a state may identify core indicators that represent the most direct measures of the WQS as the first tier of data used to support WQS attainment decisions and identify impaired waters. If measurements of these core indicators show attainment, the state may list the water as attaining the WQS. Regardless of the approach, the state should clearly document how attainment decisions are made. If not documented elsewhere, the consolidated assessment and listing method is the appropriate place.

Supplemental indicators are added to the monitoring and data collection strategy as appropriate. For example, supplemental indicators may be added for waters where there is a reasonable potential for specific pollutants to cause or contribute to water quality impairments based on evaluation of watershed conditions, including land use and source assessments. Additionally, a state may add supplemental indicators to explore the presence of pollutants widely distributed by atmospheric deposition or to establish a baseline for emerging pollutants of concern. Chapter 11 provides more discussion of potential core and supplemental indicators and how this approach may be used to improve the efficiency of water quality assessments.

It is important to note that even though the use of core and supplemental indicators should make the state, territory or authorized tribe's monitoring, information collection, and decision making activities more efficient, it cannot preclude the consideration of other relevant data and information. The state, territory, or authorized tribe is obliged to consider any other data that are relevant to its WQS (and each component) when making attainment decisions. For example, if a state shares a waterbody with another state, it must consider existing and readily available data from the state that shares the waterbody. Therefore, the assessment methodology should also address how each component of the WQS will be assessed in the event the state, territory, or authorized tribe collects or receives additional data.

3.2.2 How Does the State Interpret Data from Multiple Sources To Make WQS Attainment Impairment Decisions?

This question represents another key element of a state's consolidated assessment and listing methodology. The first step involves evaluation of the monitoring results for each indicator or type of data independently. This step includes seeking data, evaluating their quality, and interpreting the results against the applicable component of the WQS. Subsequent chapters in this document describe this process for each type of data or indicator.

The second step involves looking across the multiple types of data that serve as indicators of aquatic life–based WQS and making an attainment decision for the standard. In most cases, the WQS will be attained only when all of the indicators that the state evaluates show attainment. If one or more indicators show nonattainment, the state will typically categorize the water as not attaining the aquatic life–based WQS. There are, however, exceptions to this general policy of independent applicability, as described below.

To address the possibility of conflicting results among different types of data used to assess attainment with WQS, EPA recommends that states, territories, and authorized tribes apply the policy on independent applicability as appropriate for making WQS attainment decisions. This policy was initially crafted to address development of NPDES permit discharge limits. Its use is slightly different in the context of WQS attainment decisions.

The intent of this policy is to protect against dismissing valuable information when evaluating aquatic life use attainment, particularly in detecting impairment. EPA's policy on independent application is based on the premise that any valid, representative dataset indicating an actual or projected water quality impairment should not be ignored when one is determining the appropriate action to be taken. However, EPA recognizes that there are circumstances when conflicting results should be investigated further before the attainment or nonattainment decision is made. For example, states may obtain multiple datasets of varying quality, which may influence the reliability of the assessment results.

Figure 3-2 elaborates on the use of the independent application policy in reconciling conflicting results among different datasets used to assess attainment with aquatic life–based WQS. The decision process begins in the upper left of the figure. When a state, territory, or authorized tribe has two or more types of data that do not indicate consistent attainment status, it should determine whether differences in assessment results can be attributed to differences in the quality of the datasets. For example, this may involve consideration of analytical methods, review of sampling techniques, and detailed assessment of datasets. When the differences are due to data quality issues, the independent application policy allows for resolving the differences by cleaning the data or weighing the higher quality dataset more favorably in the attainment decision.

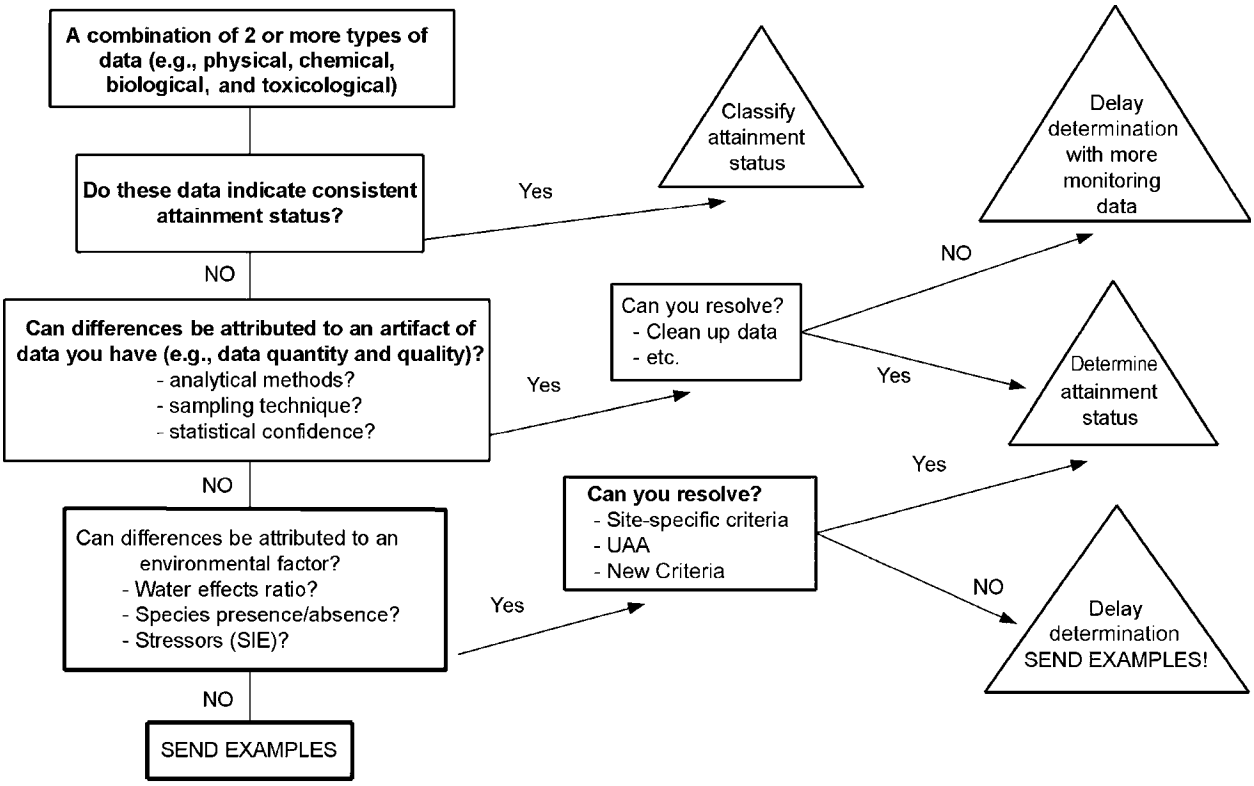


Figure 3-2. Using multiple types of data to assess attainment.

For Purposes of WQS Attainment/Nonattainment Determinations:

Policy of independent applicability *says*:

- When evaluating multiple types of data (e.g., biological, chemical) and any one type of data indicates an element of a WQS is not attained, the water should most likely be identified as impaired.
- If there is reason to doubt the nonattainment finding, re-evaluate all of the data sets to resolve discrepancies. In some cases this may lead to modification of applicable WQS to account for site-specific information.

Policy of independent applicability *does not say*:

- Always assume that a single sample result showing impairment outweighs all other data showing attainment.
- Accept all differences in data findings at face value.
- Ignore data quality and site-specific environmental factors.

When detailed data analysis fails to identify data quality issues that explain the discrepancies, site-specific environmental conditions should be considered (e.g., effects of water chemistry, or the ability of species to adapt over time). Three procedures may be explored to assess whether site-specific environmental conditions explain the discrepancies: application of the water effects ratio, development of site-specific criteria, revisions to State criteria, or conducting a use attainability analysis (UAA). These are examples of techniques that examine whether the WQS and its component elements are appropriate for the water being assessed.

Table 3-2 provides three simplified case studies demonstrating how aquatic life use support decisions are made when different types of data provide differing findings. It illustrates the importance of documenting data quality in the assessment process. EPA requests that states, territories, and authorized tribes send examples of cases in which differences in assessment results cannot be attributed either to artifacts of the data or to environmental factors. This will help the Agency further refine the independent applicability policy.

3.2.3 Examples of State Approaches To Integrate Multiple Types of Data To Assess WQS Attainment

Several States have adopted policies or legislation specifically addressing how the state defines and assesses attainment with aquatic life–based WQS. Examples from Montana and Idaho are presented below. Inclusion of these examples does not imply that these approaches are best for other states, territories, or authorized tribes. Rather, it serves to demonstrate some different approaches to documenting how the aquatic life–based WQS were assessed to identify attaining and nonattaining waters.

Montana State Profile: Montana Aquatic Life–Based WQS Assessment Process

Montana's process is presented here to illustrate a state approach to integrating multiple types of data to assess applicable WQS attainment. Additional details are available on the State's Department of Environmental Quality (DEQ) website at http://www.deq.state.mt.us/ppa/mdm.

Montana law requires the DEQ to use sufficient credible data to make WQS attainment determinations. The law defines sufficient credible data as "chemical, physical, or biological monitoring data alone or in combination with narrative information that supports a finding as to whether a waterbody is achieving compliance with applicable WQS" (75-5-103 MCA). The DEQ has developed data quality objectives to ensure that use support determinations are made with a reasonable amount of information, unless limited data provide overwhelming evidence of a water quality impairment.

The data evaluation process employs decision tables similar to the tables presented in Chapters 4 and 5 that help the reviewer score the quality of the data. For aquatic life use, the decision tables consider physical/habitat, biology, and chemistry/toxicity data. Table 3-3 presents Montana's decision table for scoring biological data for streams. Similar tables exist for physical/habitat data and chemical/toxicity data. Each category of data available for an assessment is reviewed to

Table 3-2. Applying independent applicability to cases where different data types suggest different assessment results

		Case Study 1	**Case Study 2**	**Case Study 3**
		Waterbody Description		
	Data Type	**0.8-Mile Stream Reach in Rural Watershed**	**4-Mile Coastal Blackwater Stream**	**1.5-Mile Stream Reach in Urban Watershed**
Type of Assessment Data and Information	Biological	RBP (Rapid Bioassessment Protocol) (benthic)	RBP (benthic)	RBP (benthic and fish)
	Habitat	Visually based RBP	None	None
	Toxicity	None	None	Sediment toxicity
	Physical/chemical	Conventionals	Conventionals	Conventionals and metals
Level of Information (see Table 4-2)	Biological	2	2	4
	Habitat	2	N/A	N/A
	Toxicity	N/A	N/A	4
	Physical/chemical	1	3	2
Assessment Findings	Biological	Benthos show no impairment	Benthos show no impairment	Benthos show impairment; fish show no impairment
	Habitat	Habitat shows no degradation	N/A	N/A
	Toxicity	N/A	N/A	Sediment toxicity testing indicates no exceedance
	Physical/chemical	Upstream exceedances of dissolved oxygen standard	Exceedances of pH standard	No exceedances of chronic criteria
Attainment Result	Based on decision rules documented in state, territory, or authorized tribe's assessment methodology	Attaining or inconclusive—Because of low confidence in P/C data, base decision on bio/habitat data or delay pending further monitoring.	Impaired—Investigate whether differences can be attributed to site-specific factors. If yes, develop site-specific criteria.	Impaired—Due to high confidence in biological data showing impairment.

Table 3-3. Montana's Decision Table for Scoring Biological Data (streams)

Score	Technical components	Spatial/temporal coverage	Data quality	Data currency
1	- Visual observations of biota were made with no true assessment. - Simple documentation. - Unable to make a comparison to reference condition. - Relative abundance of fish data that are not supplemented with quantitative data or cannot be interpreted by a biologist. - Fish creel surveys with limited supplemental information.	- Very limited monitoring. - Data are extrapolated from other sites.	- Data precision and sensitivity are very low or unknown. - Qualified professional does not provide any oversight. - Poor taxonomic resolution.	- Data are not relevant; biological communities may have changed significantly since the assessment was made.
2	- Only one assemblage assessed (e.g., RBP protocols). - Probable sources and causes of impairment are documented. - Reference condition can be approximated by a professional scientist. - Relative fish abundance data can be interpreted by a qualified professional or also includes quantitative fish density.	- Limited to a single sampling. - Limited sampling for site-specific studies.	- Data precision and sensitivity are low to moderate. - Data were collected following appropriate protocols; however, individuals had limited training. - Qualified professional provided oversight. - Good taxonomic resolution.	- It is unlikely that the biological communities have changed significantly since the survey was conducted.
3	- Two assemblages assessed or one assemblage with quantitative (e.g., biomass) measurements also made following standard operating procedures (SOPs). - Often include biotic index interpretations. - Fisheries data often include information about growth rates, age class, and condition; the entire fish assemblage is targeted. - Reference condition can be determined with reasonable degree of confidence and used as a basis for assessment.	- Monitoring normally occurs during a single season. - Monitoring may include site-specific studies; however, also has limited spatial coverage of the stream reach.	- Data have moderate precision and sensitivity. - Qualified professional performs survey or provides training; the individual making the survey is well trained. - Qualified professional performs the survey. - Detailed taxonomic resolution.	- Data were collected recently or it is very unlikely that the biological community has changed significantly since the survey was conducted.
4	- Two or more assemblages assessed and often include quantitative measurements following SOPs. - Reference condition is well understood and is used as the basis of the assessment. - Often include biotic index interpretations.	- Surveys conducted for multiple years and/or seasons. - Broad coverage of sites. - Often uses targeted or probabilistic design.	- High precision and sensitivity. - Assessment performed by a highly experienced qualified professional.	- Data are current; there is no doubt that the biological survey reflects current conditions.

determine its level of information, with scores ranging from a low of 1 to the highest score of 4. Scores from the different data categories are added together, and a combined score of 6 is generally considered necessary for a determination of sufficient credible data. The State does make exceptions, however, in cases where low scoring data provide overwhelming evidence (see text box) of an impairment.

Montana Criteria for Overwhelming Evidence of Impairment

Montana's methodology defines particular circumstances where data may be used to identify impairments even if the data score is less than 6. If the state, while reviewing the available data, determines that there is "overwhelming evidence" that a particular beneficial use is not supported, the use of the decision tables is unnecessary. Following are the criteria for overwhelming evidence:

- Any exceedance of an acute aquatic life standard
- A 250% exceedance of a chronic aquatic life standard, even if there is only one credible data point
- Any exceedance of an aquatic life standard based on sufficient data to calculate a geometric mean
- Any 50% exceedance of a narrative standard
- Any activities that negatively impact habitat by more than 50%
- Any activities that negatively impact biological communities by more than 50%.

Once the state has determined it has sufficient and credible data, it employs the decision criteria tables for aquatic life–based WQS for streams, lakes, or wetlands. Table 3-4 is a copy of the decision table for lakes and wetlands.

Idaho State Profile: Idaho Ecological Assessment Framework for Rivers and Small Streams

The Idaho Department of Environmental Quality (DEQ) uses a "multiple data type integration" approach to assess coldwater biota beneficial use, one of the state's aquatic life-based WQS. As part of Idaho's beneficial use reconnaissance program (BURP), the DEQ monitors a number of biological and chemical indicators. Idaho's assessment process is unique in that after considering each type of data independently to assess WQS attainment status, it combines the data into an aggregate score and uses that score as another independent measure of attainment status.

Idaho uses different bioassessment indexes for smaller streams than for larger rivers. The streams methodology is used in this example to illustrate the State's approach. Figure 3-3 demonstrates how Idaho assesses stream coldwater biota use attainment with one or more types of data. When only a single set of data exists (e.g., fish, macroinvertebrate, or physical/chemical), the DEQ applies the single data type approach illustrated on the left side of the flowchart to determine attainment of WQS. When there are two or more types of data, the DEQ uses the multiple data type integration approach illustrated on the right side of the flowchart. Idaho's multiple data type integration approach uses the following steps to determine attainment of standards for coldwater biota for streams:

Table 3-4. Montana's Decision Criteria Table for Aquatic Life-Based WQS Attainment for Lakes and Wetlands (fish, aquatic life, and wildlife)

Data category (lakes/wetlands)	Not/least impaired	Moderately impaired	Severely impaired
1. Chemistry			
Toxicity	Bioassay test indicates no acute or chronic toxicity.	Bioassay test indicates chronic toxicity.	Bioassay test indicates acute toxicity.
Chemical (toxicants, e.g., trace metals, ammonia, chlorine, organics, pesticides)[a,b] *Acute and chronic water quality standards*	For any pollutant: No exceedence of acute or chronic standard values; and/or the chronic standard values are exceeded by < 10% no more than once for one parameter in a 3-year period when measurements were taken at a minimum frequency of 4 times/year.	For any pollutant: Acute standard values are exceeded by 0.1-25%, or chronic standard values are exceeded by 0.1-50%, and/or water quality standard values are exceeded ≤10% of the measurements from a large dataset.	For any pollutant: Acute standard values are exceeded by >25%; or chronic standard values are exceeded by >50%; and/or water quality standard values are exceeded in >10% of the measurements from a large data set.
Sediment chemistry (toxicants, e.g., metals, organic compounds)	Average sediment trace metal concentrations are similar to reference condition.	Average sediment trace metal concentrations are moderately higher than reference condition.	Average sediment trace metal concentrations are substantially higher than reference condition.
Trophic status	Trophic status is similar to reference condition.	Trophic status exceeds reference condition.	Trophic status is hyper-eutrophic.
Models	Predictive models do not indicate impairment.	Predictive models indicate moderate impairment.	Predictive models indicate severe impairment.
Bioaccumulation	Pollutants are not bioaccumulated above background levels.	Bioaccumulation of pollutant is slightly above background levels.	Bioaccumulation of pollutant is substantially higher than background levels.

[a] When possible, use the average concentration of samples collected over a 96-hour period and compare directly with chronic standard values; one data point (n=1) is sufficient if no other data were collected within 96 hours.

[b] Reference condition may use a combination of the following: (1) least-impaired lake or wetland, (2) historical data, (3) upstream/down stream, (4) paired watershed, (5) review of existing literature, or (6) expert opinion.

[c] For this guidance document, exceedence is defined as a result that is higher or lower than what Montana's WQS allow.

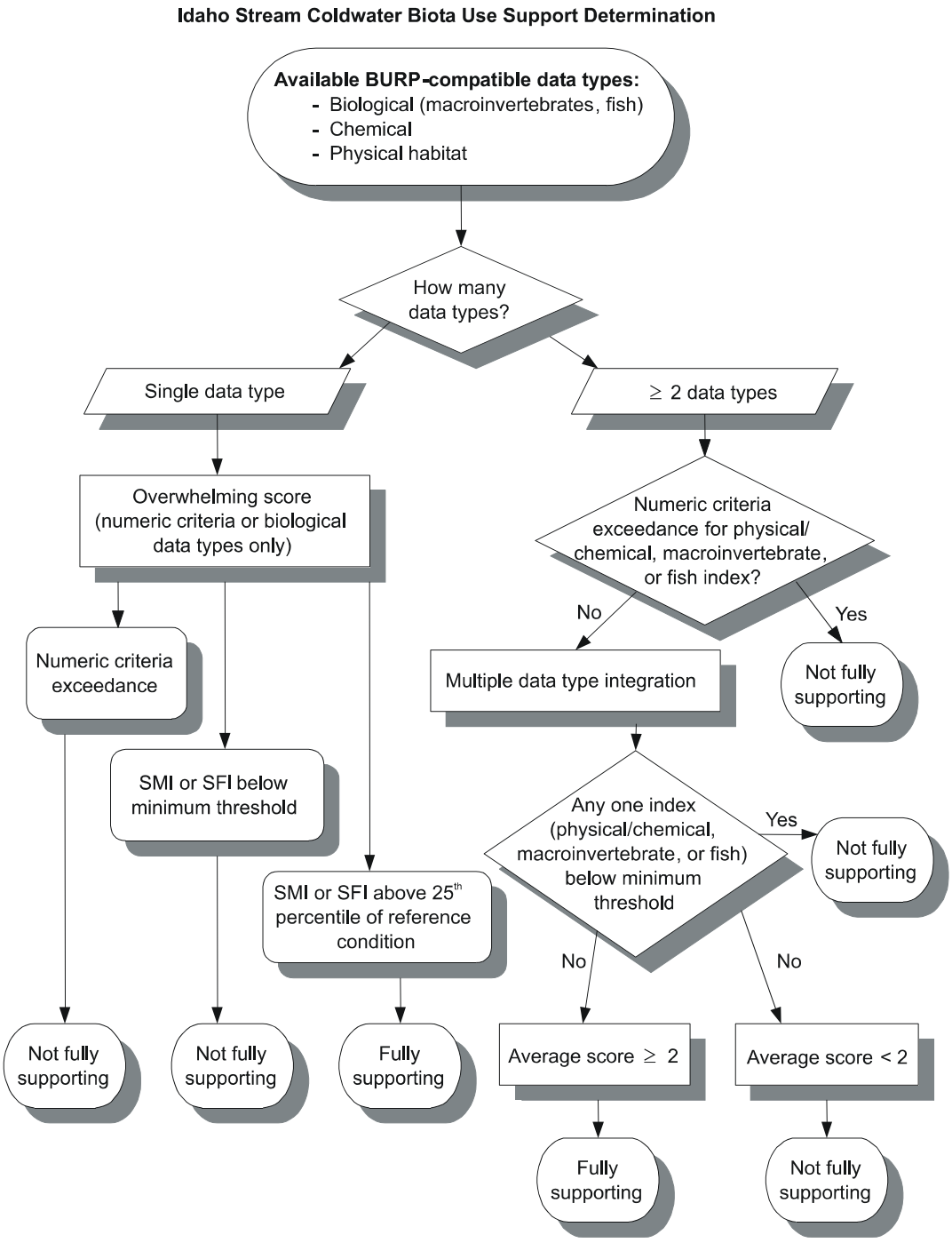


Figure 3-3. Idaho's use support determination process for stream coldwater biota use.

- Identify any numeric criteria exceedance using the criteria exceedance policy. *If there is a numeric criteria exceedance, the DEQ automatically determines the waterbody is not supporting.*

- Calculate the index scores and determine corresponding percentile categories.

- Identify any stream macroinvertebrate index (SMI) or stream fish index (SFI) scores below minimum threshold levels. *If the SMI and/or SFI scores are below minimum threshold levels, the DEQ automatically determines the waterbody is not fully supporting.*

- Identify a corresponding 1, 2, or 3 condition rating for each index. The stream habitat index (SHI) receives a 1 or 3 rating. Note that the SHI is incorporated into the combined, multi-index score, but is not considered robust enough to be used as an independent indicator of attainment status.

- Average the index ratings to determine the use support. To average the individual index ratings sum the ratings, and divide by the number of indexes uses. *An average score ≥2 is considered fully supporting. An average score <2 is considered not fully supporting.*

3.3 Recreation-Based Water Quality Standards

As discussed at the beginning of this chapter, section 101(a)(2) of the CWA establishes a goal of "fishable/swimmable" uses wherever attainable. States, territories, and authorized tribes adopt WQS to ensure that waters meet the swimmable goal. These water quality standards comprise three distinct elements: (1) designated use, (2) numeric and narrative criteria that protect the use, and (3) antidegradation policy. The form of these standards varies from state to state. Some states designate all waters for primary contact recreational use and adopt criteria to protect that use. Others assign subcategories or tiers of designated uses that reflect the nature and intensity of the use, for example, bathing beach or noncontact recreation, and criteria appropriate for each use tier. A more detailed description of subcategories of recreational uses is provided in *Implementation Guidance for Ambient Water Quality Criteria for Bacteria* (U.S. EPA 2002 - projected).

EPA's section 304(a) water quality criteria guidance for the protection of human health recommends adopting water quality criteria for two bacteria indicators for the protection of recreational uses, as appropriate. The bacteria indicators are enterococcus bacteria (for fresh or marine waters) and/or *Escherichia coli* (*E. coli*) (for fresh waters only). Many states, territories, and authorized tribes are still using the less reliable fecal coliform indicator as water quality criteria for protection of recreational uses. EPA continues to encourage states, territories, and authorized tribes that have not adopted the recommendations set forth in Ambient Water Quality Criteria for Bacteria – 1986 or other water quality criteria for bacteria based on scientifically defensible methods into their WQS to replace water quality criteria for total or fecal coliforms with criteria for *E. coli* and/or enterococci, as appropriate. In addition, the BEACH Act of 2000 amended the CWA to include section 303(i) to require states with coastal recreational waters to

adopt by April 10, 2004, WQS for pathogens and bacteria for which EPA has published criteria under CWA section 304(a).

3.3.1 What Types of Data and Information Does the State Use To Assess Whether the Recreational-Based WQS Are Attained?

For each recreation-based WQS, the state, territory, or authorized tribe should describe how it assesses attainment with the standard, and each component element. Ideally, this description should be included in the approved WQS. Alternatively, it may be defined in other implementing regulations or policies and procedures documents such as the state's continuous planning process or consolidated assessment and listing methodology.

As was described previously, states should describe the indicators and thresholds that are used to assess attainment status for each WQS. Attainment decisions for recreation-based WQS are typically based on bacteria criteria monitoring data, including the enterococci and *E. coli* indicators and fecal coliform. States, territories, and authorized tribes also consider esthetic conditions, chemical water quality criteria for protection of public health, and information on use restrictions (e.g., beach closures or public advisories). Following are brief descriptions of the various indicators or types of data that should be used to interpret recreation-based WQS.

- *Bacteria criteria*—Bacteria of fecal origin have been used for many years as an indicator of the possible presence of pathogens in surface waters and the risk of disease based on epidemiological evidence of gastrointestinal disorders from ingestion of contaminated surface water or raw shellfish. Contact with contaminated water can also lead to ear or skin infections, and inhalation of contaminated water can cause respiratory diseases. The pathogens responsible for these diseases can be bacteria, viruses, protozoans, fungi, or parasites that live in the gastrointestinal tract and are shed in the feces of warm-blooded animals.

  However, because of the difficulties in analyzing for and detecting the many possible pathogens or parasites, concentrations of fecal bacteria, including fecal coliforms, enterococci, and *E. coli*, are used as the primary indicators of fecal contamination. The latter two indicators have a higher degree of association with outbreaks of certain diseases than do fecal coliforms and were recommended as the basis for bacterial WQS in EPA's 1986 Ambient Water Quality Criteria for Bacteria document (enterococci for marine waters, *E. coli* and enterococci for fresh waters). The water quality criteria are defined as a concentration of the indicator above which the health risk from waterborne disease is unacceptably high. In 2002, EPA will publish *Implementation Guidance for Ambient Water Quality Criteria for Bacteria*, to assist states, territories, and authorized tribes in adopting and implementing these water quality criteria for recreational waters. (U.S. EPA 2002 - projected).

  See Chapter 7 for a discussion of approaches for using bacteria criteria data to assess waterbodies for recreational uses.

- *Information on indicators of fecal pollution*—Many state, territory, and authorized tribe recreation-based WQS contain narrative water quality criteria to protect waters designated for recreational use from objectionable levels of turbidity, algae, taste and odor, oil and grease, and solid waste (e.g., trash, medical waste). For example, the State of New Jersey has a narrative water quality criterion for all applicable designated uses that prohibits taste- and odor-producing substances at a level that would render the water unsuitable for the use. Data or field observations of objectionable conditions can be used to make an attainment decision, and procedures to do so should be described in the consolidated assessment and listing methodology.

  Although this document does not include a separate chapter addressing implementation of esthetic criteria, it is important that the state, territory, or authorized tribe describe how it collects and interprets information to determine attainment or nonattainment with these criteria. This description may already be developed as a translator policy or implementation procedure for that element of the WQS. If not, it should be included in the consolidated assessment and listing methodology.

- *Use restrictions and closures*—Many states, territories, and authorized tribes' water quality programs use information on bathing area restrictions and closures to determine attainment with recreation-based WQS. This information comes from state and local health departments and may be based on water quality monitoring, calibrated rainfall alert curves, or precautionary information. Before using this information on use restrictions and closures, it is important to document the basis for them. For example, the water quality agency may want to verify that the health department uses indicators and thresholds that are consistent with the state, territory, or authorized tribe's WQS.

  In general, water quality–based bathing closures or restrictions that are consistent with the state, territory, or authorized tribe's assessment methodology and are in effect during the reporting period should be used as an indicator of nonattainment. There are some exceptions, however. Bathing areas subject to precautionary administrative closures such as automatic closures after storm events of a certain intensity may not trigger an impairment decision if they are not associated with an exceedance of applicable WQS. Similarly, closures or restrictions based on other conditions like rip-tide or sharks should not trigger a nonattainment decision.

- *Chemical data*—Most recreation-based WQS include numeric chemical human health criteria for other pollutants or stressors to protect recreational uses. As noted by the Intergovernmental Task Force on Monitoring (ITFM), potentially hazardous chemicals in water and bottom sediment can be important indicators for recreational use support determinations. See Chapter 4 for discussion of chemical data.

The types of data and information available for use as indicators of recreation-based WQS attainment status come from a variety of sources. These sources range from local health departments, to interstate water resource commissions, to Federal agencies like the U.S. Geological Survey. EPA encourages states, territories, and authorized tribes to partner with

these potential sources of information when determining which indicators are most appropriate for assessing WQS attainment and when developing monitoring designs to collect appropriate data and information. For example, if a state shares a waterbody with another state, data from the state that shares the waterbody will be useful in making appropriate attainment decisions. These partnerships may reduce the monitoring and data analysis burden facing water quality assessment programs.

The assessment methodology prepared by the state, territory, or authorized tribe should describe the indicators or types of data used to assess WQS attainment status and the thresholds that distinguish attainment from nonattainment. The methodology should also describe how the state will collect and evaluate the data.

3.3.2 How Does the State Interpret Multiple Types of Data To Assess WQS Attainment?

This question represents another key element of a state's consolidated assessment and listing methodology. The first step in answering this question involves describing how the state, territory, or authorized tribe evaluates the data and information obtained for each indicator or type of data independently. This step includes seeking data, evaluating their quality, and interpreting the results against the applicable component of the state's WQS. Subsequent chapters in this document describe this process for each type of data or indicator.

The second step involves looking across the multiple types of data that serve as indicators of recreation-based WQS and making an attainment decision for the standard. In most cases, the WQS will be attained only when all of the indicators that the state evaluates show attainment. If one or more indicator show nonattainment, the state will typically categorize the water as not attaining the aquatic life–based WQS. There are, however, exceptions to this general policy, as described in Section 3.1.2.

3.3.3 Examples of State Approaches To Assess Recreation-Based WQS Attainment

EPA plans to profile different types of approaches that states, territories, and authorized tribes use to integrate multiple types of indicators or data to assess attainment with recreation-based WQS based on the documentation provided in the assessment and listing methodologies that states, territories, and authorized tribes include with the 2002 Integrated Report submissions.

3.4 Public Water Supply–Based Water Quality Standards

Public water supplies are protected under both the Safe Drinking Water Act (SDWA) and the CWA. The SDWA established a multiple-barrier approach to protecting public water supplies. The multiple-barrier approach includes assessing and protecting drinking water sources, protecting wells and collection systems, making sure water is treated by qualified operators, ensuring the integrity of distribution systems, and making information available to the public on the quality of their drinking water.

A central element of the multiple-barrier approach is the source water assessment, which is primarily a vulnerability assessment and action plan to protect source water from contamination. It includes delineation of the hydrologic boundaries of the source water and identification of potential sources of contamination within those boundaries. It then focuses on activities to prevent those sources from contaminating source waters. It does not typically include monitoring source water quality, although it may include collection and evaluation of existing water quality data. A more detailed discussion of the elements of a source water assessment is provided in the State Source Water Assessment and Protection Programs Guidance (EPA 816-R-97-009).

Under the CWA, states, territories, and authorized tribes adopt WQS for surface waters, and in some cases ground water, that protect public water supply uses. As with all WQS, public water supply–based WQS include designated use, numeric and narrative criteria to protect the use, and antidegradation policies. Water quality standards adopted into state, territory, or authorized tribal law or regulation serve to implement the SDWA's "multiple barrier" approach to drinking water protection.

To improve consistency among implementation of the SDWA and CWA, states, territories, and authorized tribes should review WQS to ensure that they have been adopted for waters delineated under source water assessments. When a state, territory, or authorized tribe has adopted WQS for the protection of drinking water uses, including public water supplies, and classified waters under its jurisdiction for such uses, those uses must be maintained and protected, consistent with section 303(c) of the CWA and the implementing Federal WQS regulations at 40 CFR 131. This includes protection of existing uses of waters as drinking water or public water supplies.

3.4.1 Which Types of Data and Information Does the State Use To Assess Whether the Public Water Supply–Based WQS Are Attained?

When adopting public water supply–based WQS, the state, territory, or authorized tribe should describe how it will assess attainment with the standard and each component element. This description should identify the indicators and thresholds that are used to assess attainment with the WQS. If this description is not a part of the approved WQS, it should be defined in other implementing regulations or policies such as the continuing planning process document or the consolidated assessment and listing methodology.

Monitoring programs should consult with the source water assessments to help identify water quality indicators that should be monitored. Following is a brief description of the various indicators or types of data that are commonly used to interpret attainment with public water supply–based WQS. More information on selecting water quality indicators is included in Chapter 11.

- *Chemical data*—To ensure that water quality protects public drinking water uses, states, territories, and authorized tribes adopt human health–based chemical criteria for waters. EPA has published section 304(a) water quality criteria guidance for the protection of

human health. Human health–based water quality criteria protect human health from exposure to carcinogens and noncarcinogenic toxicants through the consumption of drinking water and fish. Chemicals addressed through human health criteria include metals, organics, chloride, and dissolved solids. Refer to EPA's website at http://www.epa.gov/ost/humanhealth/ for more information about human health water quality criteria. Chapter 4 provides more information on using chemical data to interpret WQS attainment or nonattainment status.

Data on source water quality may be available from a variety of sources including the drinking water utilities that may screen source water before treating it. If data on source water quality are not available, states, territories, and authorized tribes may choose to evaluate monitoring data from treated or finished water supplies. These data are typically collected by the drinking water utilities to determine compliance with SDWA National Primary Drinking Water Regulations (NPDWRs or primary standards). These standards regulate the quality of treated or finished water supplied by public water systems. Primary standards protect drinking water quality by limiting the levels of specific pollutants that can adversely affect public health and are known or anticipated to occur in public water systems. Pollutants monitored for the protection of human health under the NPDWRs include volatile organic compounds, semivolatile organic compounds, inorganic constituents, salinity, radioactive constituents, and disinfection by-products.

If routine drinking water treatment is not likely to remove or alter the form or concentration of certain pollutants, states, territories, and authorized tribes may use data related to these pollutants from treated or finished water quality source water for making attainment or nonattainment decisions. In this case, levels of these pollutants in treated waters are likely to represent levels in untreated source waters. On the other hand, data on treated or finished water should probably not be used as an indicator of source water quality for pollutants that are likely to be altered, introduced to, or removed from the finished water during treatment or distribution.

- *Bacteria criteria*—Many states, territories, and authorized tribes have microbiological thresholds to protect drinking water. The NPDWRs contain criteria for treated water quality for Cryptosporidium, Giardia lambia, Legionella, total coliform, and viruses. While EPA is working on WQS for bacteria indicators to protect public water supply uses, existing standards apply to the quality of finished water rather than source water quality. However, the state, territory, or authorized tribe may choose to use these data on finished water to determine attainment with a public water supply–based WQS.

- *Use restrictions*—Another source of information that has been used by states, territories, and authorized tribes for determining whether waters attain public water supply–based WQS is drinking water use restrictions. Use restrictions include the following:

  - Closures, based on water quality concerns, of source waters that are used for drinking water supply

- Contamination-based drinking water supply advisories lasting more than 30 days per year
- Public water supplies requiring more than conventional treatment (i.e., other than coagulation, sedimentation, disinfection, and conventional filtration) due to known or suspected source water quality problems
- Public water supplies requiring increased monitoring due to confirmed detections of one or more pollutants (excluding cases with minimum detection limit issues).

The types of data and information available for use as indicators of public water supply–based WQS attainment status come from a variety of sources. These sources range from local health departments, to interstate water resource commissions, to Federal agencies like the U.S. Geological Survey. For example, many utilities collect source water quality data for process control monitoring and to comply with the Unregulated Contaminant Monitoring Rule. EPA encourages states, territories, and authorized tribes to request that drinking water utilities provide these data whenever available. EPA encourages states, territories, and authorized tribes to partner with these potential sources of information when determining which indicators are most appropriate for assessing WQS attainment and when developing monitoring designs to collect appropriate data and information. For example, if a state shares a waterbody with another state, data from the state that shares the waterbody may be useful in making appropriate attainment decisions. These partnerships may reduce the monitoring and data analysis burden facing water quality assessment programs.

3.4.2 How Does the State Interpret Multiple Types of Data To Assess WQS Attainment?

This question represents another key element of a state, territory, or authorized tribe's consolidated assessment and listing methodology. The methodology should include a description of procedures to seek data, evaluate its quality, and interpret the results against the applicable component of the WQS for purposes of making an attainment or nonattainment decision. Subsequent chapters in this document describe this process for each type of indicator or data.

All existing and readily available data and information that are consistent with the state, territory, or authorized tribe's assessment methodology must be assembled and evaluated when making a WQS attainment determination. For example, if a state shares a waterbody with another state, it must consider existing and readily available data from the state that shares the waterbody.

For purposes of making WQS attainment/impairment decisions about source waters serving as public water supplies, EPA recommends that states, territories, and authorized tribes first evaluate source water quality data. States should also evaluate treated or finished water quality data to identify potential problems with source water supplies. If one source of data indicates impairment but others do not, the reviewer should investigate the quality, quantity, and relevance of data and site-specific conditions. For example, untreated source water quality monitoring may indicate no exceedances of applicable WQS although treated water quality exceeds

maximum contaminant levels for lead and chlorine. These exceedances could be due to problems in the water treatment facility or distribution system. In this case, the source water could be assessed as attaining applicable WQS because the other data sources are not representative of the source water supply.

3.4.3 Examples of State Approaches To Assess Public Water Supply–Based WQS Attainment

EPA plans to profile different types of approaches that states, territories, and authorized tribes use to integrate multiple types of indicators or data to assess attainment with public water supply–based WQS based on the documentation provided in the assessment and listing methodologies that states, territories, and authorized tribes include with the 2002 Integrated Report submissions.

3.5 Fish and Shellfish Consumption–Based Water Quality Standards

Along with aquatic life use, fish and shellfish consumption uses compose the "fishable" goal of the CWA. Fish and shellfish consumption designated uses provide for the protection of human health related to consumption of fish and shellfish. "Fishable" means that fish and shellfish can not only thrive in a waterbody, but also, when caught, be safely eaten by humans. Although some states, territories, and authorized tribes address consumption of fish and shellfish in their aquatic life–based standards, others have a specific fish and shellfish consumption–based WQS.

3.5.1 What Type of Data and Information Does the State Use To Assess Whether Fish and Shellfish Consumption–Based WQS Are Attained?

Describing the data and information used to assess attainment with WQS is a key element of the state, territory, or authorized tribe's assessment methodology.

When adopting a fish and shellfish consumption–based WQS, the state, territory, or authorized tribe should describe how it will assess attainment with the standard and each component element. This description should identify the indicators and thresholds that are used to assess attainment with the WQS. If this description is not a part of the approved WQS, it should be defined in other implementing regulations or policies such as the continuing planning process document or the consolidated assessment and listing methodology.

Following is a brief description of the types of data and information that should be used to assess attainment with fish and shellfish consumption–based WQS.

- *Chemical data*—Three types of chemical data are used by states, territories, and authorized tribes to assess whether a particular waterbody attains fish consumption use standards: fish tissue, water column, and sediment. Chapter 4 provides more information on using chemical data to interpret WQS attainment or nonattainment status. The majority of states, territories, and authorized tribes directly monitor the level of chemical pollutants in fish and

shellfish tissue samples; however, several states monitor the level of chemical pollutants in water column and/or sediment samples.

- *Tissue data*—There are several advantages to measuring the levels of chemical pollutants directly in fish tissue samples. First, pollutant concentrations in the water column may fluctuate greatly over time. These changes occur in response to changes in chemical discharges from point and nonpoint sources as well as from fluctuations in river flow. Bioaccumulation processes occurring in the fish and shellfish act to concentrate up to 10^6 times minute levels of chemical contaminants present in the water column. In addition, levels of chemical pollutants in fish tissue tend to reflect an integration of the wide fluctuations that can occur in chemical concentrations in the water column over time. Direct measurement of the levels of chemical pollutants in fish tissues can also be used directly by states in their risk assessment methodology for calculating human health screening values and ultimately for determining fish consumption limits.

- *Water column data*—States, territories, and authorized tribes adopt human health–based chemical criteria to ensure that water quality protects fish and shellfish consumption uses. EPA has published section 304(a) water quality criteria guidance for the protection of human health. Human health–based water quality criteria protect human health from exposure to carcinogens and noncarcinogenic toxicants through the consumption of drinking water and fish. Chemicals addressed through human health criteria include metals, organics, chloride, and dissolved solids. Refer to EPA's website at http://www.epa.gov/ost/humanhealth/ for more information about human health water quality criteria.

- *Sediment data*—For some chemical contaminants that are metabolized by physiological processes in fish tissues (such as PAHs), analysis of sediment concentrations may provide a more accurate picture of the levels of environmental contamination that may result in WQS impairment. However, chemical cleanup of sediment samples prior to analysis may be both more time consuming and more expensive than direct analysis of chemical residues in fish tissue samples.

- *Fish consumption advisory information*—Fish consumption advisories are typically administered by state, territory, and authorized tribal health or environmental agencies. For information on the National Listing of Fish and Wildlife Advisories, visit EPA's website at http://www.epa.gov/ost/fish. Additional detail on how fish and shellfish consumption advisories are used to assess WQS attainment is provided under Section 3.5.2.

- *Shellfish growing area classifications*—Shellfish growing area classifications developed by the National Shellfish Sanitation Program (NSSP) can be used as part of the determinations of attainment of applicable shellfish WQS. The NSSP uses water column and tissue data (where available), along with information from sanitary surveys of the contributing watershed, to determine classifications. Certain NSSP classifications are *not* appropriate to consider when performing a beneficial use assessment. These instances are: "Prohibited" classifications set as a precautionary measure due to the proximity of wastewater treatment

discharges or absence of a required sanitary survey. Likewise, it is not appropriate to consider short-term periods when a growing area was placed in the closed status, or instances when shellfish tissue or water column pathogen data exceeded criteria, but which were not beyond the frequency, intensity (or magnitude), and duration specified in the WQS. These exceedences may be due to, for example, storm events or non-anthropogenic loadings (e.g., wildlife whose presence is not due to human influence).

- *Bacteria criteria*—Fecal coliform is the primary indicator used by the NSSP to determine whether water quality is safe for shellfish consumption.

3.5.2 How Does the State Interpret Multiple Types of Data To Assess WQS Attainment?

This question represents another key element of a state, territory, or authorized tribe's consolidated assessment and listing methodology. The methodology should include a description of procedures to seek data, evaluate their quality, and interpret the results against the applicable component of the WQS for purposes of making an attainment or nonattainment decision.

States, territories, or authorized tribes should use all relevant data and information that are consistent with its assessment methodology to assess the fish and shellfish consumption–based WQS. For example, if a state shares a waterbody with another state, data from the state that shares the waterbody should be useful in making appropriate attainment decisions. In addition to using water column data relevant to a state, territory, or authorized tribe's human health–based chemical criteria, water quality agencies often base attainment decisions on advisories and classifications provided by other organizations responsible for ensuring that fish and shellfish are safe for human consumption. Subsequent chapters in this document describe the use of chemical and bacteria criteria for making WQS attainment decisions. The use of tissue-based fish and shellfish consumption advisories and NSSP shellfish growing area classifications for making attainment decisions is described below.

On October 24, 2000, EPA issued a policy memorandum to clarify the use of tissue-based fish and shellfish consumption advisories and the NSSP classifications in WQS attainment/impairment decisions. The recommendations of this memorandum, updated to reflect the new reporting categories in the 2002 Integrated Report memorandum, are summarized below.

For purposes of determining whether a waterbody is impaired and should be included on a section 303(d) list, a fish or shellfish consumption advisory, an NSSP classification, and the supporting data would be considered existing and readily available data and information that may demonstrate nonattainment of a section 101(a) "fishable" use when:

1. The advisory is based on site-specific fish or shellfish tissue data

2. A lower than "approved" NSSP classification is not consistent with the WQS or is based on water column data showing the WQS is not attained

3. The risk assessment parameters (e.g., toxicity, risk level, exposure duration, and consumption rate) of the advisory or classification are cumulatively equal to or less protective than those in the applicable state, territory, or authorized tribal WQS.

Some fish and shellfish consumption advisories are based on Food and Drug Administration (FDA) action levels as opposed to EPA's risk-based methodology for the protection of human health. FDA action levels are established to protect consumers of interstate shipped, commercially marketed fish and shellfish rather than fish and shellfish caught and consumed within the state. FDA action levels include nonrisk-based factors (e.g., economic impacts) in their derivation, whereas water quality criteria must protect the designated uses without regard to economic impacts. EPA has therefore concluded that FDA action levels do not provide a greater level of protection for consumers of fish and shellfish caught and consumed within the state than do human health criteria. Because tissue contamination that triggers an advisory based on FDA action levels would also trigger an advisory based on human health criteria, EPA believes that a fish or shellfish consumption advisory based on FDA action levels may also indicate that section 101(a) "fishable" uses are not attained.

EPA acknowledges that, in some cases, fish and shellfish consumption advisories or restrictions may not demonstrate that a section 101(a) "fishable" use is not being attained in an individual waterbody. For example, a state may have issued a statewide or regional warning regarding fish tissue contaminated with a bioaccumulative pollutant, on the basis of data from a subset of waterbodies that do not necessarily represent the population of waters covered by this type of protective advisory. Similarly, a state may classify shellfish-growing areas "prohibited" as a precautionary measure because of the proximity of wastewater treatment discharges or where a required sanitary survey has not been conducted. Without data or information demonstrating whether the water was attaining or not attaining, there are inconclusive data to make an attainment decision.

Tissue-based fish and shellfish consumption advisories

Figure 3-4 provides EPA's recommendations for using tissue-based fish and shellfish consumption advisories when making decisions about WQS attainment status. This flowchart illustrates the conditions under which a fish or shellfish consumption advisory, and the supporting data, may demonstrate impairment of a "fishable" use for a specific waterbody.

The decision rules recommended in Figure 3-4 should apply to all pollutants that constitute potential risks to human health, regardless of the source of the pollutant. However, for fish and shellfish advisories for "dioxin and dioxin-like compounds," EPA recommends that because of the unique risk characterization issues, listing decisions should be made on a case-by-case basis. EPA is currently evaluating the role of fish advisories as part of its overall strategy to reduce

Making 303(d) & 305(b) Listing Decisions based on Fish/Shellfish Advisories

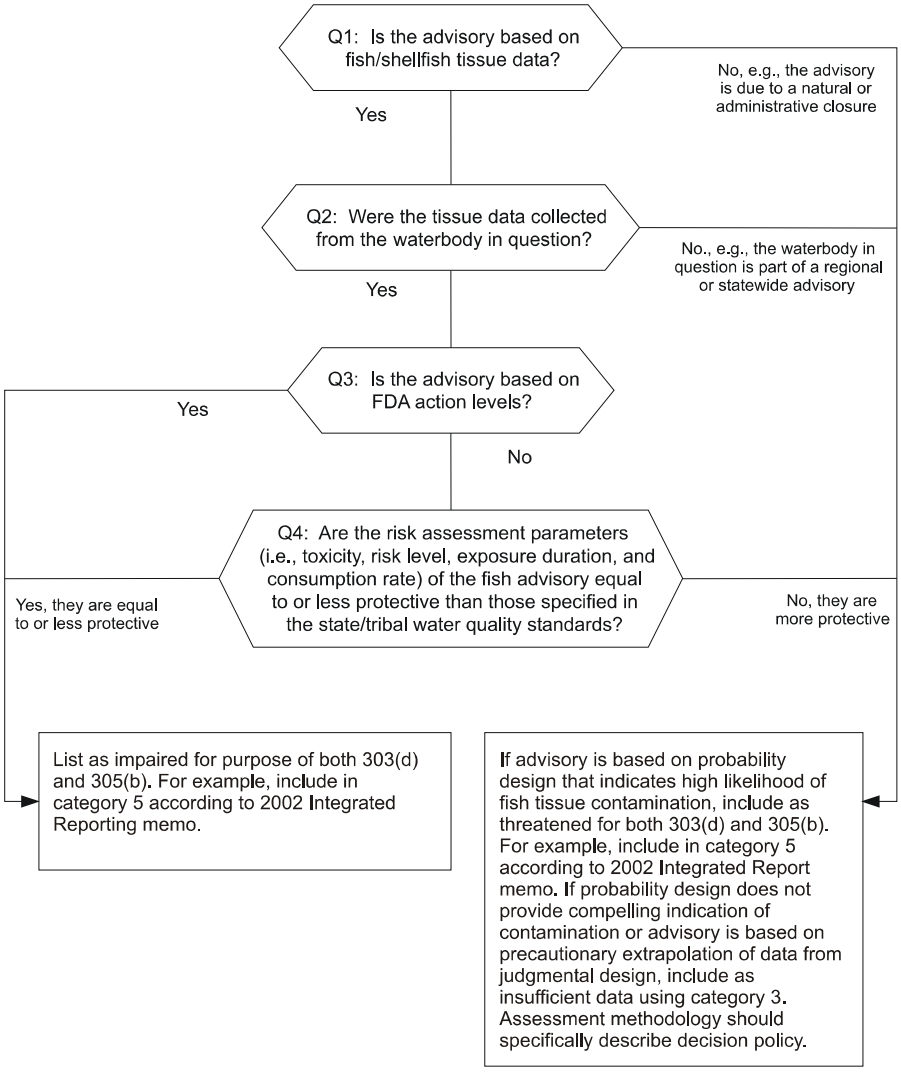


Figure 3-4. Using fish consumption advisories as indicators of WQS attainment.

human exposure to dioxin and dioxin-like compounds. EPA will be developing additional guidance specific to dioxin and dioxin-like compounds in the near future.

National Shellfish Sanitation Program classifications

The NSSP classifies shellfish-growing areas in one of five categories:

- Approved
- Conditionally approved
- Restricted
- Conditionally restricted
- Prohibited

These classifications can be used as part of the determinations of attainment of applicable shellfish WQS. The NSSP uses water column and tissue data (where available), along with information from sanitary surveys of the contributing watershed, to determine classifications. The precautionary prohibited classification is a special subcategory of prohibited that is set, without any supporting water quality data, due to the proximity of potential sources of contamination, like wastewater treatment discharges, or due to the absence of a required sanitary survey.

Before making conclusions about water quality attainment status based solely on the NSSP classifications, it is important to verify whether the WQS reflect the NSSP classification, and consider available water quality data. For example, if the state, territory, or authorized tribe's WQS specifically restricts the shellfishing use in an area that is classified as restricted, then the NSSP classification indicates the WQS attainment status of the water. In this example, the water would be identified as impaired if water quality data indicated it did not meet the definition of restricted shellfishing waters. Refer to the NSSP website for additional information on the shellfish growing area classifications.

If the WQS do not reflect the NSSP classification, the WQS could be reviewed and potentially revised to be consistent with the NSSP classification. If shellfishing is an "existing use"[1] it cannot be removed. However, if, for example, historical data and information show that the use of shellfish harvesting (or water quality that would support the use) has never been attained on a sustained basis, e.g., has had a "conditionally approved" shellfish classification, the WQS may be considered for revision to recognize the conditional nature of the shellfishing under the "subcategory of use" provision in 131.10(c). To do so, conduct a Use Attainability Analysis (UAA) using the tests in 131.10(g), and include in the description of the subcategory of use the maximum number of times per year the area would be closed or restricted from shellfish harvesting. This description should be based on historical data and model information, and should include a plan to review the WQS in accord with any remediation or long-term control plans to address the causes of the conditions resulting in the episodic nonattainment.

[1]40 CFR 131.3(e): "*Existing uses* are those uses actually attained in the water body on or after November 28, 1975, whether or not they are included in the water quality standards."

Figure 3-5 suggests an approach for using NSSP growing area classifications to assess shellfishing use support. In general, the approved and conditionally approved classifications are supporting the use, unless water quality monitoring data indicate otherwise. The restricted, conditionally restricted, and prohibited (other than precautionary prohibited) classifications should be considered impaired, unless water quality monitoring data or applicable WQS indicate otherwise.

3.5.3 Examples of State Approaches To Assess Fish and Shellfish Consumption–Based WQS Attainment

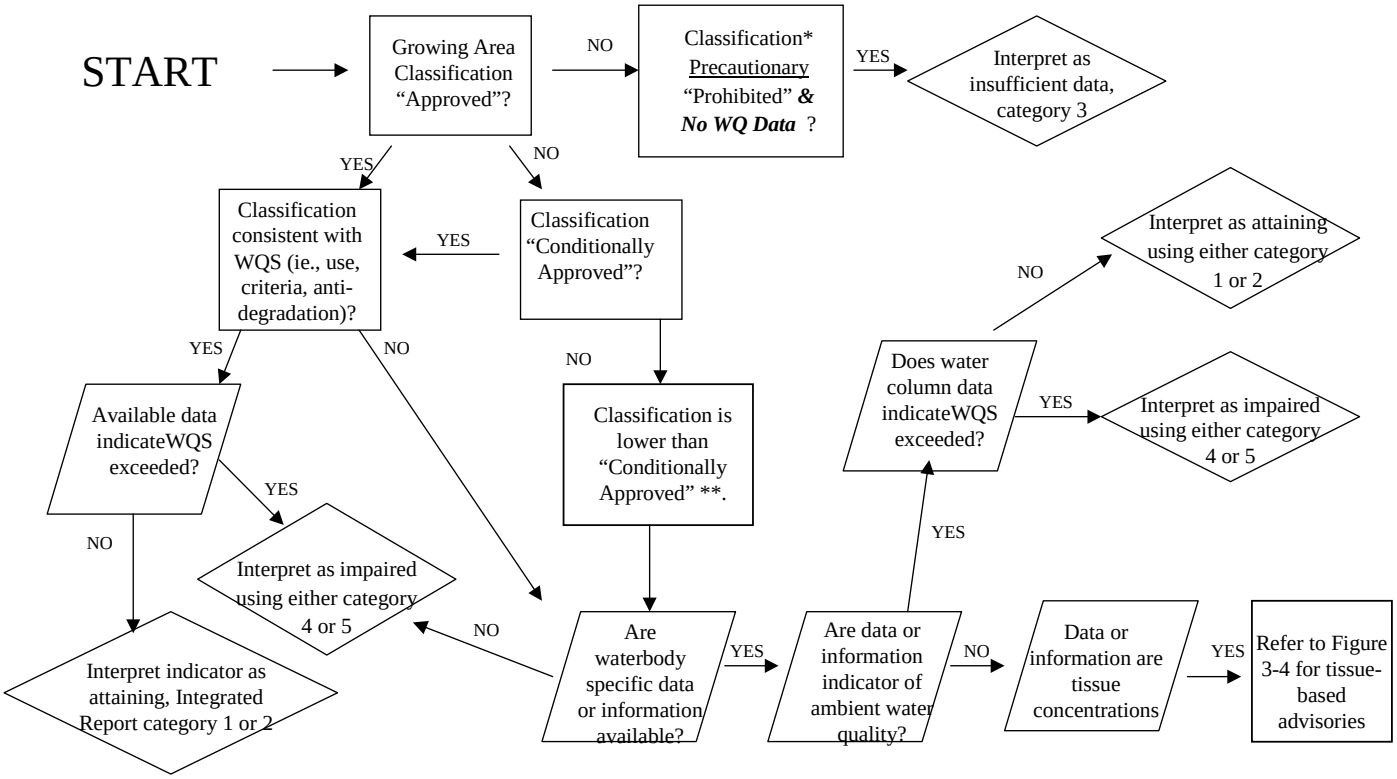
EPA plans to profile different types of approaches that states, territories, and authorized tribes use to integrate multiple types of indicators or data to assess attainment with public water supply–based WQS based on the documentation provided in the assessment and listing methodologies that states, territories, and authorized tribes include with the 2002 305(b) and 303(d) submissions.

Following are brief highlights of the programs of two states: Vermont and North Carolina.

Vermont uses fish tissue mercury data to assess fish consumption use in the state's lakes. These data are used to determine attainment of standards for toxic substances and habitat (because of accumulation up the food chain). The state assigns a finding of impairment to waterbodies only where there is a "no consumption" advisory for a subpopulation of enhanced sensitivity (with supporting fish tissue data) and where the target species are actually present. Vermont employs a probability design to determine which lakes should be sampled for mercury in fish tissue. The state recognizes that this design produces data that better represent actual mercury levels in the target population. Relying solely on consumption advisories based on limited targeted monitoring designs may give false impressions of where problems exist.

North Carolina employs two shellfish classifications, Conditionally Approved-Open and Conditionally Approved-Closed, that require interpretation regarding appropriate attainment and reporting decisions under 303(d) and 305(b). Since North Carolina's Conditionally Approved-Open appears to be equivalent to NSSP Conditionally Approved, the waterbody is in attainment if the WQS identifies shellfish harvesting as a designated use that is attainable except in certain conditions. As long as the classification and WQS are consistent and monitoring data continue to indicate they are met, the water is in attainment. North Carolina's Conditionally Approved-Closed appears to be equivalent to NSSP Conditionally Restricted or Restricted classifications. Again, unless the WQS is consistent in excluding shellfish harvesting as a use, the waterbody is not in attainment of WQS, and therefore should be so identified as impaired.

Using NSSP Growing Area Classifications as Indicator of WQS Attainment



* 1999 NSSP Model Ordinance Subsection IV.@03: Growing Area Classification

Restricted, Conditionally Restricted, or Prohibited (not precautionary prohibited)

Figure 3-5. Using NSSP growing area classifications as indicator of WQS attainment.

3.6 References

Association of State and Interstate Water Pollution Control Administrators. 1997. Important concepts and elements of an adequate state watershed monitoring and assessment program. Paper prepared for ASIWPCA Standards and Monitoring Task Force by Chris Yoder under cooperative agreement to EPA Office of Water, August 1997.

National Research Council (NRC). 2001. Assessing the TMDL Approach to water quality management. Washington DC: National Academy Press.

U.S. Environmental Protection Agency (U.S. EPA). 1991. EPA policy on independent application. Transmittal memorandum of final policy on biological assessment and criteria from Tudor Davies to EPA Regions, June 19, 1991.

U.S. EPA. 1997. State source water assessment and protection programs guidance. EPA 816-R-97-009. August 1997.

U.S. EPA. 2000. Use of fish and shellfish advisories and classifications in 303(d) and 305(b) listing decisions - WQSP-00-03. Transmittal memorandum of policy on use of fish and shellfish consumption advisories from Geoffrey Grubbs and Robert Wayland, October 24, 2000.

U.S. EPA. 2002 (projected date). Implementation guidance for ambient water quality criteria for bacteria. Office of Water, Washington, DC.

Van den Berg, et al. 1998. Toxic equivalency factors (TEFs) for PCBs, PCDDs, PCDFs for humans and wildlife. Environ Health Perspect 106(12):775–792.

4. Using Chemical Data as Indicators of Water Quality

Contents

4. Using Chemical Data as Indicators of Water Quality

Complete assessment of water quality demands consideration of different types of data because each type provides unique insights into water quality standards (WQS) attainment status. This chapter addresses the role of chemical data in assessing WQS attainment status and listing impaired waters. Subsequent chapters cover biological, toxicity, bacteria, and habitat data, respectively. Note that conventional indicators such as temperature, pH, and dissolved oxygen, which are sometimes referred to as physical data, are included in this chapter because they are generally treated as chemical indicators of water quality.

Chemical data are important indicators of WQS attainment/impairment for a number of reasons. All state, territory, and authorized tribal WQS include chemical-specific numeric water quality criteria adopted to protect aquatic life and human health from the effects of pollution. Assessments of chemical concentrations serve as direct measures of stressors to aquatic life and human health. Chemical-specific data and water quality models allow predictions of the likelihood of impacts to aquatic life and human health where they may not yet have occurred. Chemical pollutants also lend themselves to chemical-specific total maximum daily limit (TMDL) development and source controls, particularly as expressed in National Pollution Discharge Elimination System (NPDES) discharge permits.

Using chemical data involves issues related to data quality as well as ensuring that data are representative of water quality conditions. This chapter helps states reduce uncertainty by documenting their approaches for using chemical data to make WQS attainment decisions and list impaired waters. Each section title poses a question that addresses an element of a state's assessment and listing methodology.

4.1 How Are Chemical Data Used Within the Context of the State's WQS?

State WQS play a central role in a state's water quality management program. Standards drive water quality assessments, 303(d) lists of impaired waters, 305(b) reports on water quality status and trends, TMDLs, NPDES permits, and nonpoint-source management measures. These standards include designated uses appropriate for each waterbody, numeric and narrative criteria adopted to protect uses, and policies to prevent degradation of waters. Chemical data primarily support assessments of the extent to which numeric and narrative criteria are met. The state's assessment and listing methodology should describe how chemical data are collected and how they are used to determine attainment of WQS.

States, territories, and authorized tribes adopt water quality criteria to protect designated uses, including aquatic life, recreation, public water supplies, and fish and shellfish consumption. The criteria should be based on sound scientific rationale and should contain sufficient indicators or parameters to protect the designated uses. Water quality criteria are numeric criteria derived from EPA's 304(a) criteria guidance documents or other scientifically defensible methods, or narrative criteria adopted when numeric criteria cannot be determined or to supplement numeric criteria.

4.1.1 Numeric Criteria

Under section 304 of the CWA, states, territories, and authorized tribes adopt chemical-specific numeric criteria into their WQS to protect designated uses. These criteria generally include:

- Aquatic life thresholds for acute or chronic exposure of sensitive organisms,
- Human health thresholds for cancer risk or noncancer risk due to exposure via drinking water and fish tissue consumption, and
- Organoleptic effect thresholds for drinking water consumption and recreation.

The Final Water Quality Guidance for the Great Lakes System outlines a process for developing numeric criteria to protect wildlife (U.S. EPA 1995).

EPA, under section 304(a) of the CWA, periodically publishes recommendations (guidance) for use by states and authorized tribes in developing and adopting criteria protective of designated uses. In adopting such criteria, states and tribes may use (1) 304(a) criteria, (2) 304(a) criteria modified to reflect site specific conditions, or (3) other scientifically defensible methods (see 40 CFR 131.11). The complete listing of EPA-recommended water quality criteria can be found in National Recommended Water Quality Criteria–Correction (EPA 822-Z-99-001) or the most recent update thereof at http://www.epa.gov/ost/standards/wqcriteria.html. Table 4-1 lists state websites where individual state WQS, including numeric criteria, are presented in detail. A source for the effective state water quality standards and criteria is on the EPA website at: http://www.epa.gov/wqsdatabase.

Numeric criteria for aquatic life protection

Development of numeric water quality criteria for aquatic life protection is a complex process described in the 1985 Guidelines and in EPA's criteria guidance documents and summarized in the Water Quality Standards Handbook (U.S. EPA 1994). The process involves collecting and analyzing data on a specific chemical concerning its toxicity to aquatic organisms. To serve as a basis for criteria development, data must be available for at least one species in each of at least eight different families. If sufficient data are available, EPA derives a recommended acute (criteria maximum concentration, CMC) and chronic (criteria continuous concentration, CCC) criterion. Acute thresholds estimate the highest 1-hour concentration that will not have an unacceptable lethal effect on 95% of the species tested. Similarly, chronic thresholds estimate the highest 4-dayconcentration that should not cause unacceptable toxicity during long-term exposure. Acute or chronic criteria can be adjusted to reflect water quality characteristics such as pH, temperature, or hardness. Separate criteria may be developed for fresh and salt waters.

Table 4-1. State agency web sites for water quality standards and criteria (current as of February 2001)

State	WQS Web Address
AK	http://www.state.ak.us/local/akpages/ENV.CONSERV/dawq/dm/wqsmain/regs.htm
AL	http://www.adem.state.al.us/RegsPermit/ADEMRegs/Div6Vol1/rdiv6v1.html http://www.adem.state.al.us/RegsPermit/PropRules/proprule.htm
AR	http://www.adeq.state.ar.us/regs/reg02.htm
AZ	http://www.sosaz.com/public_services/Title_18/18-11.htm
CA	http://www.swrcb.ca.gov/plnspols/index.html
CO	http://www.cdphe.state.co.us/cdphereg.asp#wqreg
CT	http://dep.state.ct.us/wtr/wqs.pdf
DE	http://www.dnrec.state.De.us/water/wqs1999.pdf
FL	http://www.dep.state.fl.us/ogc/documents/rules/shared/62-302.pdf http://www.dep.state.fl.us/ogc/documents/rules/shared/62-302t.pdf
GA	http://www.ganet.org/dnr/environ/rules_files/exist_files/391-3-6.pdf
HI	http://mano.icsd.hawaii.gov/doh/rules/ADMRULES.html
IA	http://web.legis.state.ia.us/Rules/2000/iac/567iac/56761/
ID	http://www2.state.id.us/adm/adminrules/rules/IDAPA58/58INDEX.HTM http://www2.state.id.us/adm/adminrules/bulletin/sept00.pdf
IL	http://www.ipcb.state.il.us/title35/download/C302.pdf
IN	http://www.ai.org/legislative/iac/title327.html
KS	http://www.kdhe.state.ks.us/download/index.html#bowreports
KY	http://www.lrc.state.ky.us/kar/401/005/026.htm
LA	http://www.deq.state.la.us/planning/regs/title33/index.htm#partix
MA	http://www.state.ma.us/dep/brp/wm/files/314cmr4.pdf
MD	http://209.15.49.5/dsd_web/default.htm
ME	http://janus.state.me.us/legis/statutes/38/title38ch30sec0.html
MI	http://www.deq.state.mi.us/swq/
MN	http://www.revisor.leg.state.mn.us/arule/7050/
MO	http://mosl.sos.state.mo.us/csr/10csr/10c20-7.pdf
MS	http://www.deq.state.ms.us/newweb/opchome.nsf/pages/SurfaceWaterfiles/$file/wqc.pdf
MT	http://www.deq.state.mt.us/dir/Legal/Chapters/CH30-06.pdf
NC	http://mapsweb01.sips.state.nc.us/ncoah/ncadministrativ_/title15aenviron_/chapter02enviro_/default.htm
ND	N/A
NE	http://www.deq.state.ne.us/RuleAndR.nsf/pages/117-TOC
NH	http://www.des.state.nh.us/wmb/Env-Ws1700.pdf
NJ	http://www.state.nj.us/dep/landuse/njac/7-9b.pdf http://www.state.nj.us/dep/watershedmgt/swqs/

Table 4-1. State agency web sites for water quality standards and criteria (current as of February 2001) (continued)

State	WQS Web Address
NM	http://www.nmenv.state.nm.us/NMED_regs/swqb/20nmac6_1.html
NV	http://www.leg.state.nv.us/NAC/NAC-445A.html
NY	http://www.dec.state.ny.us/website/regs/ch10.htm
OH	http://www.epa.state.oh.us/dsw/rules/3745-1.html
OK	http://www.state.ok.us/~orwb/rules/Chap45.pdf
OR	http://waterquality.deq.state.or.us/wq/wqrules/wqrules.htm
PA	http://www.pacode.com/secure/data/025/chapter93/chap93toc.html
RI	http://www.state.ri.us/dem/REGS/WATER/QUALREGS.PDF
SC	http://www.scdhec.net/water
SD	http://legis.state.sd.us/rules/rules/7451.htm
TN	http://www.state.tn.us/sos/rules/1200/1200-04/1200-04.htm
TX	http://www.tnrcc.state.tx.us/oprd/rules/pdflib/307`.pdf
UT	http://www.rules.state.ut.us/publicat/code/r317/r317-002.htm
VA	http://ftp.deq.state.va.us/pub/watrregs/wqs.zip
VT	http://www.state.vt.us/wtrboard/july2000wqs.htm
WA	http://www.ecy.wa.gov/biblio/wac173201a.html
WI	http://www.legis.state.wi.us/rsb/code/nr/nr100.html
WV	http://www.state.wv.us/csr/docs/WPDocs/4601 .wpd

Note: N/A means WQS not on the web or web address not available at time of compilation.

EPA's criteria guidelines for aquatic life protection recommend that a criterion is comprised of a chemical concentration, a duration, and a frequency. The acute criterion (criteria maximum concentration, CMC) equals the highest concentration of a pollutant to which the aquatic species can be exposed for a short period of time without deleterious effects. The chronic criterion (criteria continuous concentration, CCC) equals the highest concentration of a pollutant to which the aquatic species can be exposed for an extended period of time (4 days) without deleterious effects. For ammonia, a 30-day rather than a 4-day average is recommended. Alternative averaging periods can be developed by using data that relate toxic response with exposure time, or by using models of toxicant uptake and action (U.S. EPA 1991b). Both the acute and chronic exposure durations were set to be fully protective of fast-acting toxicants, and are therefore even more protective for slower acting toxicants.

Early in the WQS program, EPA criteria guidance for several parameters, including chlorides, turbidity, and temperature, stated that these criteria should not be exceeded at any frequency. Later EPA guidance distinguished between conventional pollutants and toxic pollutants when providing recommendations about the number of exceedances that constitute nonattainment of WQS. For conventional pollutants, the 305(b) guidelines indicated that whenever more than 10% of the water quality samples collected exceed the criterion threshold, the WQS is not attained (U.S. EPA 1997).

EPA recommended that acute and chronic aquatic life criteria for toxics not be exceeded more than once every 3-year period on the average. EPA selected this frequency to provide a level of protection similar to the 7Q10 design flow or low-flow condition. The exceedance frequency recommendation is considered protective. Like the magnitude and duration components of the water quality criteria, it may also be revised to reflect site-specific information on exposure and response relationships.

Numeric criteria for human health protection

States adopt numeric chemical criteria for human health protection as part of WQS developed to protect public water supply, fish consumption, and recreational uses of surface waters. States may adopt numeric fish-tissue-based chemical criteria for the protection of human health from consumption of contaminants such as mercury in fish.

In 2000, EPA published revisions to the methodology for developing water quality criteria for the protection of human health (U.S. EPA 2000c). These revisions incorporate the latest scientific information for developing water quality criteria, including systematic procedures for evaluating cancer risk, noncancer health effects, human exposure, and bioaccumulation potential in fish. (See also http://www.epa.gov/ost/humanhealth/method/index.html.)

The revised methodology provides more flexibility for decision making at state, tribal, and EPA regional levels. Specifically, it provides opportunities for states, territories, and tribes to use tailored information on fish consumption rates, acceptable risk levels, and other factors that influence the calculations of chemical criteria. EPA believes that adoption of water quality

criteria requires several risk management decisions that are often better made at the state, territory, or tribal level.

Water quality criteria to protect human health are generally based on protecting against long-term exposure to low concentrations of a toxic pollutant. When a chemical human health criterion is applied to WQS attainment decisions, EPA recommends evaluating comparing the mean (or geometric mean if appropriate for a skewed data set) of the measured concentrations with the criterion. However, some states have adopted human-health-based chemical criteria that establish instantaneous maximum concentrations, for which any exceedance constitutes nonattainment. If the mean or geometric mean exceeds the criterion, the WQS is not being attained.

4.1.2 Narrative Criteria

To supplement numeric criteria for toxic chemicals, states adopt narrative criteria. These criteria help ensure that all designated uses are protected under a wide range of circumstances. Narrative criteria are effective tools for addressing toxic effects of pollutants, exposure pathways, or exposure conditions for which the state has not adopted chemical-specific numeric criteria. Recommended narrative criteria, which are often referred to as "free froms," were first developed in 1968 and continue to be an important element of state, territorial, and tribal WQS.

EPA guidance explains that these "free froms" should apply to all waters of the United States at all flow conditions, including ephemeral and intermittent streams (U.S. EPA 1994). Narrative criteria guidance indicates that all waters should be free from substances that:

- Cause injury to, are toxic to, or produce adverse physiological responses in humans, animals, or plants
- Settle to form objectionable deposits
- Float as debris, scum, oil, or other material in concentrations that form nuisances
- Produce objectionable color, odor, taste, or turbidity
- Produce undesirable aquatic life or result in the dominance of nuisance species

States, territories, and authorized tribes may use chemical data are used to interpret a narrative criterion. For example, a state may use chemical concentrations in sediment, in conjunction with other information on sediment toxicity and the health of benthic communities, to identify a water as impaired because of sediment contamination. Another example is the use of fish tissue data. The concentrations of pollutants in fish tissue can be used in risk-based calculations to assess attainment of the fish consumption use as well as to issue fish consumption advisories. States may use narrative criteria to determine that a surface water is impaired for its public water supply use. This decision might be triggered by a finding that a drinking water utility has violated a chemical-specific maximum contaminant level for treated water and that the chemical is present in the surface water.

EPA recently issued guidance on the use of fish and shellfish consumption advisories and certain shellfish growing area classifications in determining attainment of water quality standards and listing impaired waterbodies under section 303(d) of the Clean Water Act. This guidance also recommends (in part) that in instances where tissue concentrations of pollutants do not indicate an exceedance, states, territories, and authorized tribes translate the applicable narrative criteria on a site-specific basis or adopt site specific numeric criteria to account for higher than expected exposures from contaminated fish or shellfish tissue and protect designated uses. This is discussed in an October 24, 2000, letter issued by EPA's Office of Science and Technology and Office of Wetlands, Oceans and Watersheds (U.S. EPA 2000d).

EPA encourages states, territories, and authorized tribes to use chemical data to interpret narrative criteria; however, these jurisdictions should develop implementation procedures, often referred to as translators, that explain how different types of chemical data are used to make attainment/impairment decisions based on narrative criteria. These implementation procedures should be made available for review and comment by the public.

4.2 What Actions Does the State Take To Assess and Document Data Quality, Including Third-Party Data?

This is an important question because it acknowledges that all data may not be of equal value for assessing WQS attainment/impairment. Results of analyses of chemical data or any other type of data may be of limited value unless they are accompanied by documentation about sample collection, analytical methods, and quality control (QC) protocols. Poorly documented monitoring results may indicate potential problems, corroborate other data and information, or trigger additional monitoring, but they may not to provide an adequate basis of an attainment or impairment decision if they fail to meet accepted data quality requirements. Chemical data with good data-quality documentation should be used to support an attainment/impairment decision.

Several states are reexamining and better defining requirements for acceptable data and protocols for screening data adequacy prior to interpreting data to make WQS attainment decisions. EPA has extensive technical documents on this topic, some of which are listed in the references to this chapter. Documenting data quality requirements and data evaluation procedures is a critical element that states need to address in their WQS implementation procedures.

It is important to balance data quality requirements with common sense. Data quality requirements must be objective and inclusive. States, territories, and authorized tribes must consider all existing and readily available data when making WQS attainment/impairment decisions. For example, if a state shares a waterbody with another state, it must consider existing and readily available data from the state that shares the waterbody. Data should not be excluded solely because of their source or their age, without a reasonable explanation as to why they do not represent water quality conditions. Similarly, data collected using methods different from those the state prefers should be considered if the detection limits for the method are appropriate for both the criteria threshold and the concentration detected.

Whatever data collection methods are employed, the state should ensure that their sampling design program be implemented in such a way that each sample represents the variable conditions extant in the target water(s) at the locality, during the time period for which the WQS attainment/impairment assessment is intended. Probability-based sampling designs are one way of ensuring representativeness of the sample, and permit estimation of the uncertainty associated with the sample-based estimates of the means, proportions, and other statistics required for comparison to criterion values. When these conditions are met, procedures for formal statistical inference can be applied to make scientifically defensible WQS decisions (see Appendices C and D; Peterson et al. 1999).

4.2.1 How Does the State Define Data Quality?

EPA encourages states, territories, interstate commissions, and authorized tribes to use the data quality objectives process to define minimum data quality requirements. This includes information on appropriate sample size and monitoring design, sample collection and handling protocols, analytical methods and detection limits, QC procedures, and data management. Frequently this type of information is documented in the state's quality assurance (QA) project plan or standard operating procedures (SOPs) for monitoring. Data quality requirements may be defined in the applicable WQS or in other implementating regulations or policy and procedures documents. For example, the WQS may define critical conditions, such as flow or temperature, under which the criteria apply or should be modified, while the implementation procedures may discuss information like data quality objectives, samples sizes, and SOPs.

It is important to make this information available to other organizations such as tribal, interstate, state, Federal, academic, and volunteer citizen groups that also monitor water quality. Over time, these potential partners may agree to meet data quality requirements if an agency clearly spells out these requirements in its assessment and listing methodology or other readily available and well-publicized documents.

Sample size is an important element of data quality. In general, statistical tests have good power for detecting exceedances if they are based on data from samples composed of 30 or more sampling units. Smaller sample sizes are prone to yield erroneous attainment decisions because they have a low probability of detecting WQS exceedances unless they are large and pervasive. Figure 4-1 illustrates the effect of sample size on the probability of detecting exceedances when the actual proportion of exceedant sampling units exceeds the criterion proportion (e.g., 10% exceedance) by amounts between zero and 50%. Computing such power curves for different sample sizes can be an effective tool for illustrating the benefits of larger sample sizes. Similarly, computing the width of the confidence intervals for the exceedance proportion provides an indication of the uncertainty of the sample estimate for distinguishing impairment from attainment (Figure 4-2). Appendices C and D describes these issues in more detail and provides guidance and additional references on determining sample sizes.

If a state is uncomfortable basing attainment/impairment decisions on small data sets, it should demonstrate a commitment to collection of sufficient data to support its data quality objectives.

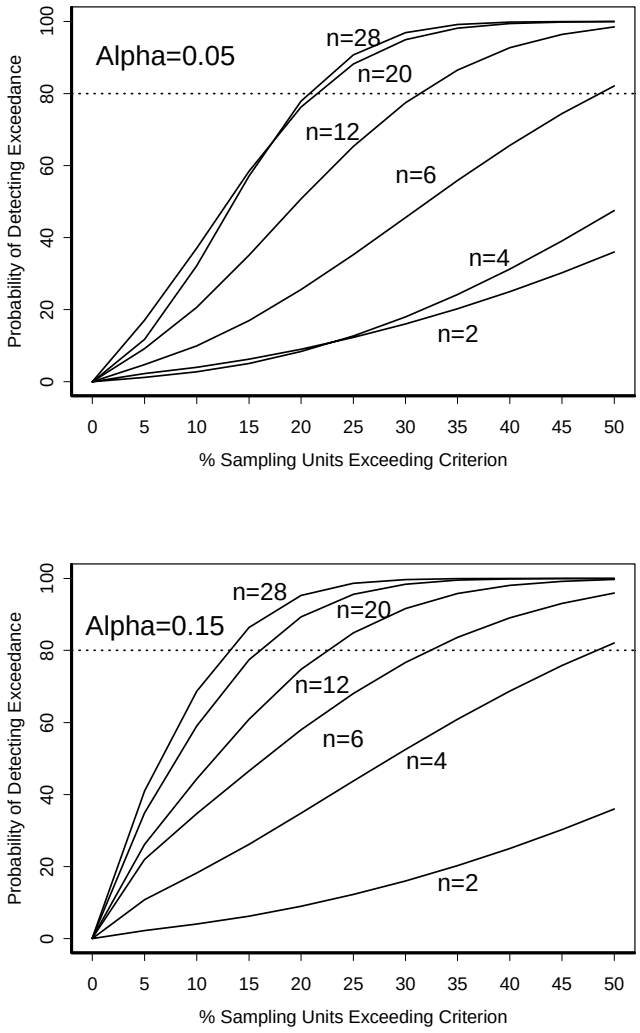


Figure 4-1. Power curves computed from different sample sizes, at two different alpha levels. In the upper panel, the Type I error rates (α) are held at ≤0.05, while those in the lower panel are held at α ≤0.15. The vertical axis is the probability of detecting that a sample exceeds the criterion exceedance rate, and the horizontal axis is the percent that a particular sample actually exceeds the criterion exceedance rate. Statisticians call the vertical and horizontal axes variables "power" and "detectable effect size," respectively. A horizontal reference line at 80% in each graph marks what many researchers consider the minimally acceptable power level for a statistical test. It is noteworthy that when α ≤0.05, none of the sample sizes shown has acceptable power for detecting exceedances ≤20%. However, when the acceptable Type I error is increased to α=0.15, sample sizes of 28 and 20 have sufficient power for detecting exceedances between 13% and 20%. Smith et al. (2001) recommend that WQS tests should have sufficient power to detect exccedances of ≥15% above criterion. The complex relationships among sample size, effect size, Type I error rates (α), and Power (1-Type II error rate) are discussed in detail in Appendices C and D.

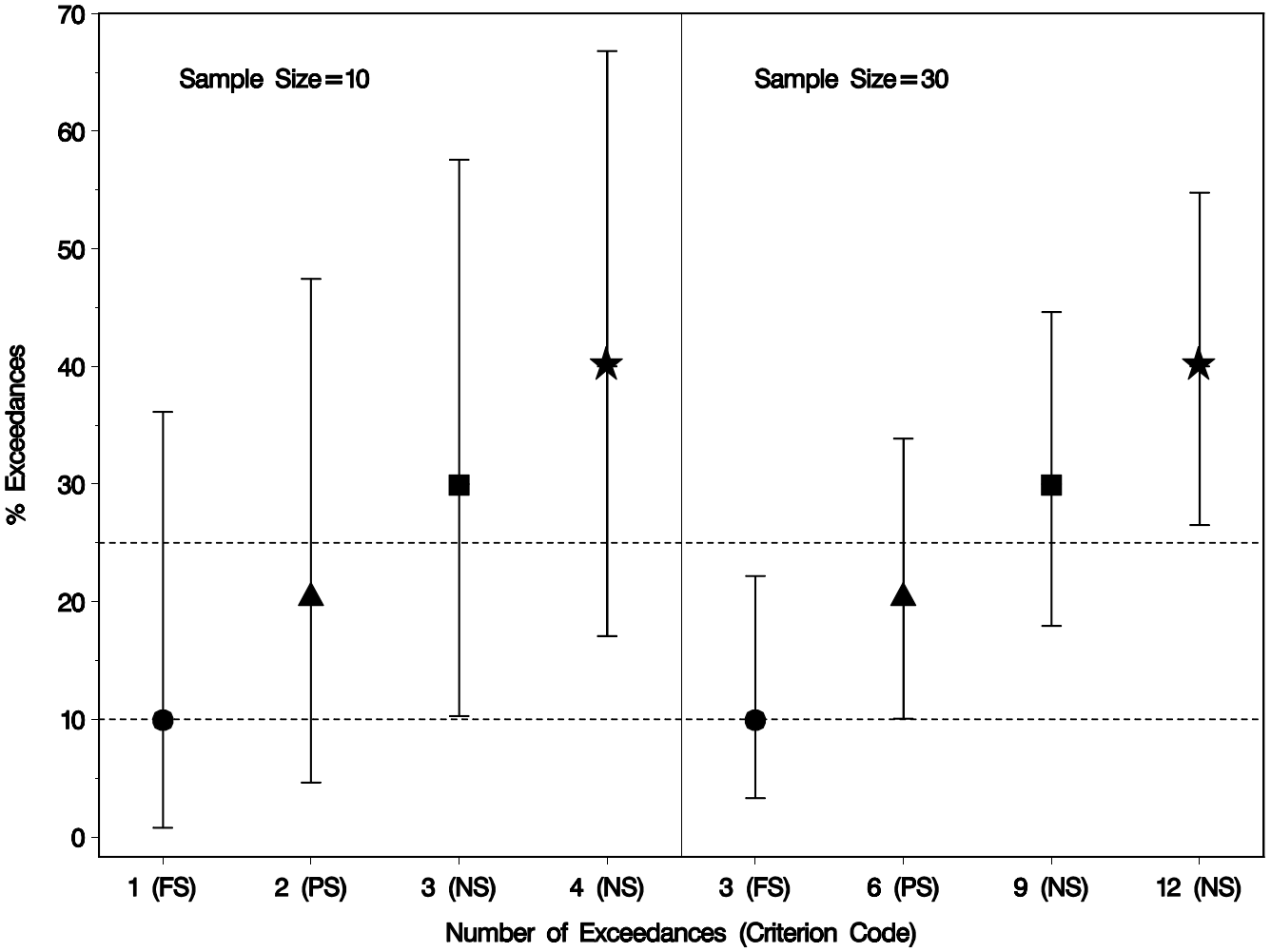


Figure 4-2. Effect of sample size on confidence interval widths. Two sets of two-sided 85% confidence intervals are illustrated (the 85% confidence level is recommended in Appendices C and D). For each sample size, confidence limits were estimated for samples with 10% (dot), 20% (triangle), 30% (square), and 40% (star) exceedances among their constituent sampling units. The two horizontal lines denote the 10% exceedance criterion and the 25% threshold (criterion value + 15% minimum detectable effect size) recommended by Smith et al. (2001). Exceedances of ≤10% constitute full support (FS) of the the 10% criterion; 10%-20% exceedances constitute partial support (PS); and exceedances >20% imply no support (NS). Any two-sided confidence interval whose lower bound extends below the 10% reference line satisfies the criterion. Thus with a sample size of 10, a minimum of 30% exceedances (= 3 sampling units) are required, compared with only 20% exceedances (= 6 sampling units) for a sample of 30 sampling units.

4.2.2 How Does the State Assess (Review and Evaluate) Data Quality?

The term "data quality assessment" means the scientific and statistical evaluation of data to determine whether data obtained from monitoring operations are of the right type, quality, and quantity to support water quality attainment decisions. Data quality does not exist in a vacuum; one must know in what context a data set is to be used in order to determine whether it is adequate.

Figure 4-1 shows the probability of detecting very low exceedances with small data sets, unless the actual rate of exceedance in the water is very high or common. Larger data sets have a greater probability of detecting less frequent exceedances. If a small data set detects an exceedance, the waterbody is likely experiencing a higher frequency of exceedances. However, if a small data set does not detect an exceedance, it is difficult to say with statistical confidence that the water is attaining WQS. Larger data sets are more powerful in terms of supporting decisions that a water is attaining WQS.

Guidance for assessing the quality of available data sets is provided in *Practical Methods for Data Quality Assessment* (EPA/600/R-96/084). For assessing WQS attainment, EPA recommends a tiered approach. The following steps should be part of the first tier of a data quality review process:

- Screen documentation to determine whether appropriate procedures were used and QA/QC measures were in place (e.g., if the third party's field and laboratory procedures are documented in SOPs).

- Determine whether samples were collected under the appropriate conditions for comparison with WQS (e.g., correct time of year or flow conditions).

- Review sample collection and analytical methods to determine compatibility with the state agency's QA/QC requirements and SOPs; also determine if the third party's sample collection and analytical methods were actually followed in creation of the data set.

- Determine whether the metadata accompanying the data set meet the agency's requirements (e.g., determine adequacy and accuracy of geographic documentation in the data set).

If the data do not meet all of the above conditions, they may be rejected from further analysis. Once it is determined that the data set meets the state's basic documentation requirements, the evaluators might decide to do additional screening of the actual data sets. At a minimum, they might want to look for values below the detection limit of the analytical method, because these may influence how the data set is analyzed or incorporated with other data. If, upon analyzing the data, evaluators suspect errors in the collection or analysis, they may want to conduct more in-depth analysis of QA/QC procedures. This screening could include reviews of QA/QC reports to determine if the data set meets the agency's QA/QC requirements regarding documenting measurement system performance (e.g., adequate use of QC samples), the approach to handling missing data and nondetects, and deviations from SOPs.

4.2.3 How Does the State Document the Level of Data Quality?

The 305(b) Consistency Workgroup developed a table assigning qualitative levels of information or data quality to different types of chemical data. Several states have since developed similar approaches for rating the quality of data used in WQS assessments. States are encouraged to use an approach similar to that described in Table 3-2 to report on the quality of data supporting attainment/impairment decisions. In addition, they should begin documenting quantitative information about the quality of these decisions.

The data hierarchy described in Table 4-2 addresses data quality considerations such as sample collection and analytical techniques, spatial and temporal representativeness, and QA procedures. The user rates the data set on the basis of the rigor of the information, where 1 is the lowest and 4 is the highest. In general, Level 1 information alone may not be sufficient for an attainment decision; however, even a short period of record can indicate impairment in cases of gross exceedances of criteria.

States should supplement the data descriptions illustrated in Table 4-2 with more quantitative descriptions of the confidence and power of their attainment/impairment decisions. This documentation clearly illustrates to decision makers and the public the impact of small data sets on the uncertainty in the water quality decision. Quantitative documentation of the uncertainty is expressed in statistical terms of error rates, both Type I decision error, or the α-level, and Type II decision error, or the β-level, of the assessment. These decision errors are discussed in detail in Appendices C and D. A Type I error occurs when an attaining waterbody is erroneously judged to be impaired, and a Type II error occurs when an impaired waterbody is erroneously judged to be attaining. EPA encourages states to collect sufficient numbers of samples to balance both types of error at reasonable levels.

To summarize, for attainment decisions based on chemical data, states should document:

- Level of information based on Table 4-2 or state-developed table or approach;
- Sample size, range of concentrations, mean, median, and standard deviation; and
- Level of statistical confidence (Type I decision error and Type II error) and width of the confidence interval.

4.3 How Does the State Analyze and Interpret Chemical Data to Determine WQS Attainment/Impairment?

The most important element of the state, territory, or authorized tribe's assessment and listing methodology is documentation of how the state analyzes data to determine WQS attainment and identify impaired waters. This documentation should be consistent with the state's, territory's, or authorized tribe's implementation procedures that are described either in the WQS or alternatively in other implementing regulations or policies and procedures documents. If the implementation procedures do not describe how WQS should be interpreted for determining

Table 4-2. Hierarchy of physical/chemical data levels for evaluation of aquatic life use attainment

Level of info[a]	Sample collection and analytical techniques	Spatial and temporal representativeness	Data quality
1	Any one of the following: • Water quality monitoring using grab water sampling • Water data extrapolated from an upstream or downstream station where homogeneous conditions are expected • Best professional judgment based on land use data, source locations	Low spatial and temporal coverage: • Quarterly or less frequent sampling with limited period of record (e.g., 1 day) • Limited data during key periods or at high or low flows (critical hydrological regimes)[b] • Data are >5 years old and are not reflective of current conditions	Approved QA/QC protocols not followed or QA/QC results inadequate Methods not documented Inadequate metadata
2	Any one of the following: • Water quality monitoring using grab water sampling • Rotating basin surveys involving multiple visits or automatic sampling • Synthesis of existing or historical information on fish contamination levels • Screening models based on loadings data (not calibrated or verified)	Moderate spatial and temporal coverage: • Bimonthly or quarterly sampling during key periods (e.g., spring/summer months) • Fish spawning seasons, including limited water quality data at high and low flows • Short period of record over a period of days or multiple visits during a year or season • Data are <5 years old and there is high certainty that conditions have not changed since sampling	Approved SOPs used for field and lab; limited training Low precision and sensitivity QA/QC protocols followed; QA/QC results adequate Adequate metadata
3	Any one of the following: • Composite or a series of grab water sampling used (diurnal coverage as appropriate) • Rotating basin surveys involving multiple visits or automatic sampling • Calibrated models (calibration data <5 years old)	Broad spatial and temporal (long-term, e.g., > 3 years) coverage of site with sufficient frequency and pollutant coverage to capture acute events: • Typically, monthly sampling during key periods (e.g., spring/ summer months, fish spawning seasons), multiple samples at high and low flows • Lengthy period of record (sampling over a period of months) • Data are <5 years old and there is high degree of certainty that conditions have not changed since sampling	Moderate precision and sensitivity Samplers well trained SOPs used for field and lab Moderate precision/ sensitivity QA/QC protocols followed; QA/QC results adequate Adequate metadata
4	Follows defined sampling plan which includes the following elements: • Description of how sample is representative of target populaiton • Defined data quality objectives, including error rate, confidence interval, sample size	Broad spatial (several sites) and temporal (long-term, e.g., > 3 years) coverage of site with sufficient frequency and coverage to capture acute events, chronic conditions, and all other potential chemical impacts • Monthly sampling during key periods (e.g., spring/summer months • Fish spawning seasons (including multiple samples at high and low flows) • Continuous monitoring • Data are <5 years old and there is high degree of certainty that conditions have not changed since sampling	High precision and sensitivity Samplers well trained SOPs used in field and lab QA/QC protocols followed; QA/QC results adequate Adequate metadata

[a]Level of information refers to rigor of chemical sampling and analysis, where 1 = lowest and 4 = highest.

[b]Even a short period of record can indicate a high confidence of *impairment* based on chemical data; 3 years of data are not required to demonstrate impairment. For example, a single visit to a stream with severe acid mine drainage impacts (high metals, low pH) can result in high confidence of impairment. However, long-term monitoring may be needed to establish full attainment.

attainment status, the procedures should be revised to, at a minimum, reference the assessment and listing methodology.

In recent years, most water quality agencies have followed approaches developed by the 305(b) Consistency Workgroup for interpreting data to assess WQS attainment/impairment status as described in the 305(b) reporting guidelines (U.S. EPA 1997). Guidance documents for developing section 303(d) lists of impaired waters indicate that waters identified as partially or not supporting WQS according to the 305(b) guidelines should be included on 303(d) lists of impaired waters (U.S. EPA 1991a). Under the Integrated Report, waters that are impaired or threatened for one or more designated uses can belong in Category 4 or Category 5 and may or may not require a TMDL depending on the source of impairment and the management action that has been completed on the waterbody. The Integrated Report addresses one area where the 305(b) and 303(d) guidances differed, the treatment of waters that are "fully supporting water quality standards, but threatened." The 305(b) guidance had a broader definition of waters fitting this category than 303(d) did, so it was not appropriate to assume that all threatened waters in 305(b) reports belonged on 303(d) lists. The Integrated Report clarifies there are only three instances in which waters that are threatened for one or more designated uses do not require the development of a TMDL: 1) if a TMDL has been completed, 2) if other pollution control requirements are reasonably expected to result in attainment in the near future, or 3) if the water is threatened by something other than a pollutant. In each case, follow-up monitoring should be scheduled for these waters to verify attainment of the WQS as expected. Table 4-3 reflects these modifications to the decision rules found in the 305(b) guidance made to simplify the reporting categories and to clarify the linkages between 303(d) lists of impaired waters and 305(b) water quality inventory reports. This table reflects the Integrated Report, therefore, does not include the "fully supporting, but threatened" category. The table also combines the "partially" and "not supporting" categories into a single category called "impaired."

An assessment methodology should take into account the balance between desired data requirements and the practical realities affecting the availability of information and the strength of the available evidence. For example, a state's methodology could describe an acceptable probability of decision errors for making an attainment decision, except in cases where overwhelming evidence of impairment is found. Examples of overwhelming evidence could be a single sampling event showing dangerously low pH downstream of an abandoned mine, fish kills that cannot be attributed to natural causes, elevated levels of accumulative pollutants in fish tissue. Another example could be allowing the results from analytical methods with high detection levels or poor sensitivity (e.g., field test kits) in cases where the results clearly suggest large exceedances of criteria. Photographs or other documentation of gross impairment may also be considered, if appropriate.

Generally, decisions should be based on very small data sets only when there is overwhelming evidence for impairment. EPA does not recommend making decisions based on small data sets of water column chemistry for attainment. Therefore, in the overwhelming majority of WQS scenarios, an approach based on probability sampling, in which states define an acceptable probability of decision error, will be preferred. Statistical inference based on sequential sampling designs may offer an alternative that allows states use defined data quality objectives to

Table 4-3. Interpreting chemical data to assess WQS attainment

Type of criteria	Attaining WQS	Impaired for 305(b) and 303(d)	Example of statistical guidelines for documenting data quality objectives for attainment decisions
Acute chemical criteria for toxic pollutant for the protection of aquatic life	For any one pollutant, no more than one excursion above acute criterion (EPA's criteria maximum concentration [CMC] or applicable state/tribal criterion) within a 3-year period based on grab or composite samples	More than one excursion above criterion within any 3-year period	One-sided binomial confidence intervals on the percentage of samples whose hourly mean exceeds the stated acute 1-hour mean criterion value. Type I and Type II error rates should be approximately equal to 0.15 and the minimum effect size set at 15% (0.15). The tests and confidence intervals evaluate: H_0: ≤5% of the samples exceed the 1-hour acute criterion value H_a: >5% of the samples exceed the 1-hour acute criterion value
Chronic chemical criteria for toxic pollutant for the protection of aquatic life	For any pollutant, no more than one excursion above chronic criterion (EPA's criteria continuous concentration [CCC] or applicable state/tribal criterion) within a 3-year period based on grab or composite samples.	More than one excursion above criterion within any 3-year period	One-sided binomial confidence intervals on the percentage of samples whose 4-day mean exceeds the stated chronic 4-day mean criterion value. Type I and Type II error rates should be approximately equal to 0.15 and the minimum effect size should be set at 15% (0.15). The tests and confidence intervals evaluate: H_0: ≤5% of the samples exceed the 4-day chronic criterion value H_a: >5% of the samples exceed the 4-day chronic criterion value.
Acute or chronic chemical criteria for conventional pollutant	For any pollutant, no more than 10% of the samples exceed the criterion	More than 10% of the samples exceed the criterion	One-sided binomial confidence intervals on the percentage of aliquots whose pollutant concentration exceeds the criterion value. Type I and Type II error should be approximately equal to 0.15 and the minimum effect size set at 15% (0.15). The tests and confidence intervals evaluate: H_0: ≤10% of the aliquots in the sample exceed the criterion value H_a: >10% of the aliquots in the sample exceed the criterion value.

Table 4-3. Interpreting chemical data to assess WQS attainment (continued)

Type of criteria	Attaining WQS	Impaired for 305(b) and 303(d)	Example of statistical guidelines for documenting data quality objectives for attainment decisions
Human health criteria for drinking water, fish consumption, recreation, or other human-health related uses	Annual mean concentration does not exceed criterion	Annual mean concentration exceeds criterion	Lower one-sided confidence intervals (or corresponding upper -one-sided t-tests) on the sample mean, geometric mean, or median pollutant concentration, pH, etc., relative to the stated criterion value. Type I and Type II error rates should be set at ≤0.15 and the minimum effect size at 15% (0.15). The tests and confidence intervals evaluate: H_0: mean, geometric mean, or median of the sample ≤ the criterion value H_a: mean, geometric mean, or median of the sample > the criterion value
Human health criteria for fish and shellfish consumption	Tissue levels do not exceed state/tribal risk-based levels, and/or water column concentrations do not exceed human health criteria	Tissue levels exceed state/tribal risk-based levels, and/or water column concentrations exceed human health criteria	Lower one-sided confidence intervals (or corresponding upper -one-sided t-tests) on the sample mean, geometric mean, or median pollutant concentration, pH, etc., relative to the stated criterion value. Type I and Type II error rates should be set at ≤0.15 and the minimum effect size at 15% (0.15). The tests and confidence intervals evaluate: H_0: mean, geometric mean, or median of the sample ≤ the criterion value H_a: mean, geometric mean, or median of the sample > the criterion value

identify impaired waters with small data sets. When a state describes its acceptable levels of decision error, it is able to identify the corresponding number of exceedances within a particular sample size that meet the level of decision error. With sequential sampling designs, the state, territory or authorized tribe may make an impairment decision once enough samples that fail to meet the WQS are collected and additional sample collection can be curtailed (see Figure 4-3).

Ecological applications of sequential sampling have been described (Carter et al. 1994); presumably this approach could be extended to many of the WQS situations described in this document.

4.3.1 What Statistical Analyses for Interpreting Chemical Data Does the State Use?

EPA acute and chronic chemical criteria for protection of aquatic life are examples of "ideal standards" as defined by Barnett and O'Hagan (1997). "Ideal standards" include criteria set as maximum levels not to be exceeded. As defined by Barnett, "ideal standards" pose several challenges in assessing attainment. The standard set as a not-to-be-exceeded chemical criterion does not address variation and uncertainty; therefore, assessing attainment implies a monitoring design that measures for the chemical throughout the entire population—all points in the waterbody continuously over time (Barnett 1997). Any state monitoring program to collect data for interpreting attainment with WQS, however, involves sampling the population and estimating the characteristics of the population on the basis of the characteristics of the sample. The use of sampling introduces variability and uncertainty. Some of this is due to the natural variability of the waterbody and human error associated with sample collection and analysis. A key element of the uncertainty relates to the precision of the sample. A larger, well-conducted monitoring effort will yield better, more precise estimates of the true condition than a smaller or poorly run effort. It is important that a state's data quality objectives and QA/QC procedures clearly define adequate statistical and other implementation procedures to ensure that all parties are aware of the minimum data set and statistical analysis requirements to show attainment (Barnett 1997). Figure 4-2 illustrates the effect of sample size on the confidence intervals and, therefore, the precision associated with attainment decisions.

Barnett recommends development of a statistically verifiable ideal standard composed of two parts. First is the ideal standard or criterion. Barnett provides several examples for defining the criterion based on a maximum, an average, or a percentile. The choice should reflect the pollutant–effect relationship. For example, pollutants that have a threshold effect should have a criterion that specifies that a high percentage of the samples must lie below the threshold in order to limit the amount of time the pollutant levels exceed the threshold (Barnett 1997).

The second component of the statistically verifiable ideal standard includes guidelines for statistical verification of the criterion (Barnett 1997). These guidelines, which should be a component of the state's WQS implementation procedures, describe a level of assurance that the criterion is attained. For example, the guidelines may describe the acceptable error rates for Type 1 and Type 2 error, the size of the confidence interval, and the sample size.

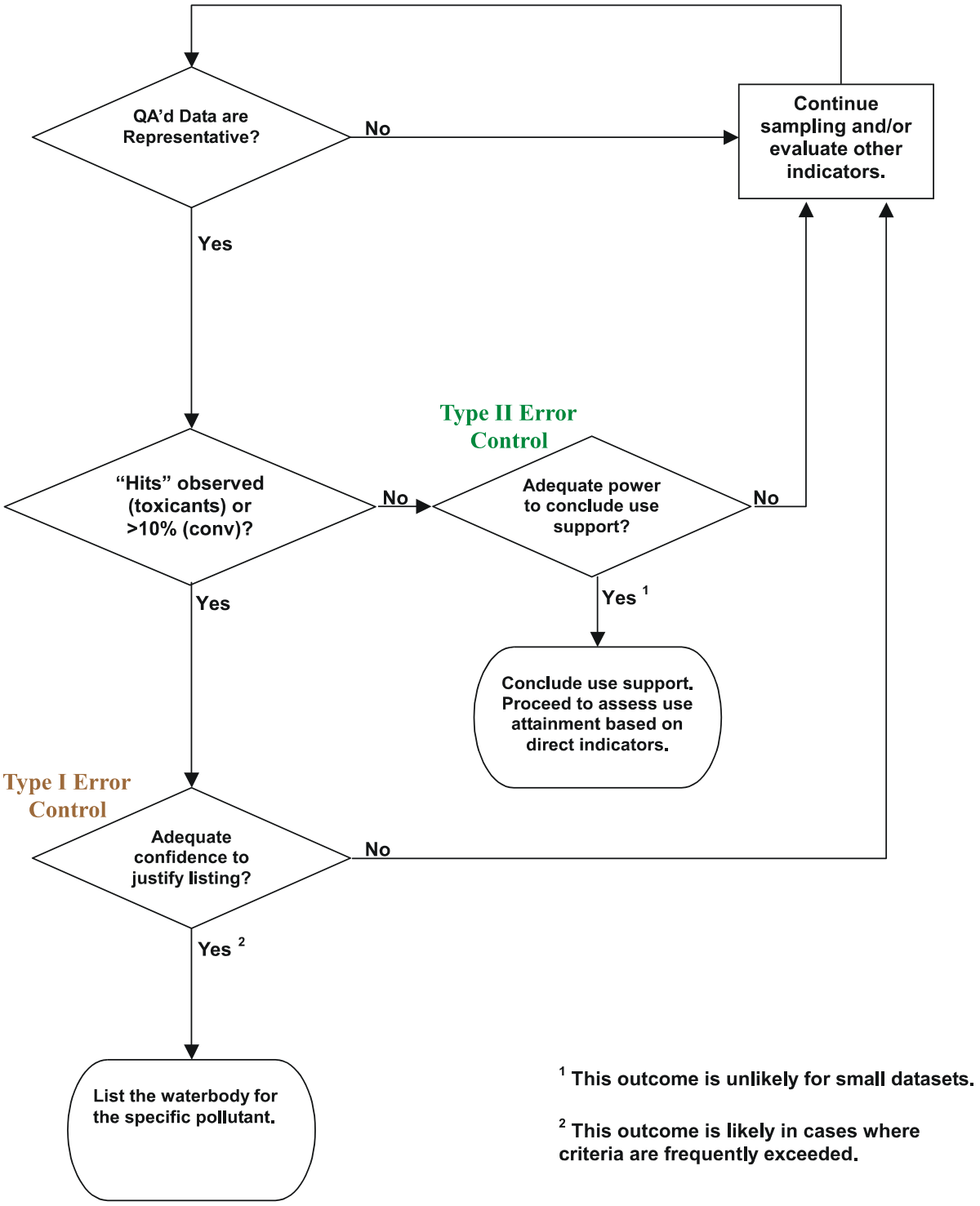


Figure 4-3. Sequential Decisionmaking. Making use support decisions, based on small sets of water column chemistry data, while balancing the risk of Type I (false positive) and Type II (false negative) decision errors.

Statistical methods are widely available to account for uncertainty and can be used to set appropriate bounds on how attainment should be demonstrated. Appendices C and D explore statistical hypothesis testing, confidence intervals, and Type I and Type II error and provides guidance on how to use parametric and nonparametric hypothesis tests to evaluate WQS attainment. The appendices include examples of the estimation of proportions of exceedances and upper concentration percentiles such as would be appropriate for comparison to threshold values. Methods for estimating mean and geometric mean concentrations and local variability and testing hypotheses about them are also included in the appendices. Tests and estimates of mean concentrations are appropriate for many of the human health criteria.

Another important analytical tool is trend analysis. Although state WQS generally call for analysis of data collected over 1- to 3-year periods in making attainment decisions, states should analyze data over longer time periods when they are available.

4.3.2 How Does the State Make Attainment/Impairment Decisions in the Absence of a "Perfect Data Set"?

State assessment and listing methodologies should describe how to make water quality attainment/impairment decisions in the absence of complete data sets that meet all their data quality requirements. For example, if a state's methodology calls for 30 samples to make an attainment decision with specified confidence, what will the state do if it collects only 10 samples? A state should develop procedures for looking for overwhelming evidence of water quality impairment, such as a single sample with well-documented QC methods that shows a large exceedance of an applicable criterion.

Another factor the state may want to consider is the effect of the sample size on the likelihood of detecting an exceedance of the criteria in the first place, particularly for chemicals that are not naturally occurring in the environment. For example, if a waterbody actually experiences 2 days during which a pollutant concentration exceeds the criterion, the probability that a sample of 36 daily samples collected over 3 years will capture both excursions is less than 1%. To detect such a small number of exceedances, one would have to collect about 600 samples. See Appendices C and D for further discussion and assumptions leading to the above statements about probabilities.

The state's methodology should also identify other types of data or information that may be used to support or supplement a sparse data set to make an attainment/impairment decision, such as the use of point-source discharge data from facilities discharge monitoring reports to predict ambient concentrations using dilution models. Regardless of the approach used, the state should clearly document how attainment decisions are made. If not documented elsewhere, the consolidated assessment and listing method is the appropriate place.

4.4 References

Barnett V, O'Hagan A. 1997. Setting environmental standards: The statistical approach to handling uncertainty and variation. London: Chapman & Hall.

Carter JL, Ravlin FW, Fleischer SW. 1994. Sequential egg mass sampling plans for gypsy moth (*Lepidoptera: Lymantriidae*) management in urban and suburban habitats. J Econ Entomol 87(4):999-1003.

Peterson SA, Urquhart NS, Welch EB. 1999. Sample representativeness: A must for reliable regional lake condition estimates. Environ Sci Technol 33:1559-1565.

Smith EP, Ye K, Hughes C, Shabman L. 2001. Statistical assessment of violations of water quality standards under Section 303(d) of the Clean Water Act. Environ Sci Technol 35:606-612.

U.S. Environmental Protection Agency (U.S. EPA). 1986. Guidelines for deriving numerical national water quality criteria for the protection of aquatic organisms and their uses. NTIS PB85-227049.

U.S. EPA. 1991a. Guidance for water quality-based decisions: The TMDL process. EPA/440/4-91-001.

U.S. EPA. 1991b. Technical support document for water quality-based toxics control. EPA/505/2-90-001.

U.S. EPA. 1994. Water quality standards handbook: 2nd ed. EPA-823-B-94-005A.

U.S. EPA. 1995. Final Water Quality Guidance for the Great Lakes System. Federal Register March 23, 1995; page 15365.

U.S. EPA. 1997. Guidelines for preparation of the comprehensive state water quality assessments (305(b) reports) and electronic updates. EPA-841-B-97-002A and 002B.

U.S. EPA. 1998. EPA guidance for quality assurance project plans. EPA/600/R-98/018.

U.S. EPA. 2000a. Guidance for data quality assessment. EPA/600/R-96/084.

U.S. EPA. 2000b. Guidance for the data quality objective process. EPA/600/R-96/055.

U.S. EPA. 2000c. Methodology for deriving ambient water quality criteria for the protection of human health. EPA-822-B-00-004.

U.S. EPA. 2000d. Use of fish and shellfish advisories and classifications in 303(d) and 305(b) listing decisions - WQSP-00-03. Transmittal memorandum of policy on use of fish and shellfish consumption advisories from Geoffrey Grubbs and Robert Wayland, October 24, 2000.

5. Using Biological Data As Indicators of Water Quality

Contents

5. Using Biological Data As Indicators of Water Quality

A primary objective of the Clean Water Act is to restore and maintain the chemical, physical, and biological integrity of the Nation's waters (Section 101(a)). In 1991, EPA issued a policy statement regarding the "Use of Biological Assessments and Criteria in the Water Quality Program" (U.S. EPA 1991a). This policy states in part:

> *To help restore and maintain the biological integrity of the Nation's waters, it is the policy of the Environmental Protection Agency that biological surveys shall be fully integrated with toxicity and chemical-specific assessment methods in State water quality programs. EPA recognizes that biological surveys should be used together with whole-effluent and ambient toxicity testing, and chemical-specific analyses to assess attainment/non-attainment of designated aquatic life uses in State water quality standards. EPA also recognizes that each of these three methods can provide a valid assessment of designated aquatic life use impairment.*

The framework described in this chapter is intended to help states and other jurisdictions make better decisions when using biological assessment data and other data to determine impairments and list waterbodies. This framework includes a discussion of the key elements a state's methodology should contain for using biological assessment data and provides information on different methodologies and approaches that can be used to support different water quality determinations.

The references section of this chapter lists key EPA guidance documents that provide technical information to develop and implement effective bioassessment programs for assessing attainment of water quality standards and identification of impaired waters. All of these documents are available through EPA's website: http://www.epa.gov/ost/biocriteria/index.html.

Throughout this chapter, the various key elements and approaches for using biological data are rated from Level 1 to Level 4, to reflect the rigor and level of quality each provides. Level 4 is the highest rigor and quality and provides a relatively high level of certainty in an assessment. Level 1 data describe much less rigorous approaches that are still valuable but present a relatively high degree of uncertainty in the assessments and decisionmaking based on those assessments.

5.1 How Does the State Use Biological Data in the Context of WQS?

Biological assessments, or bioassessments, are an evaluation of the biological condition of a waterbody using biological surveys and other direct measurements of the resident living organisms. Biological assessment data are important for measuring the attainment of WQS for the protection of aquatic life. Biological assessment data can provide a clear picture of whether a waterbody is meeting its designated aquatic life use(s) and can validate whether existing water quality criteria for toxic chemicals, whole-effluent toxicity, physical characteristics, and habitat quality are adequately protecting the designated aquatic life use(s). Biological assessments reflect the total cumulative impact of all stressors over a period of time on a waterbody on the biological community. As such, they are a unique waterbody response measurement, providing information about a waterbody that no other measurement can. For this reason, a state should

use biological assessment data as a core indicator for making aquatic life use determinations, as long as the state can provide documentation of the adequate quality and rigor of the key elements of the state's bioassessment program.

Biological data of different varieties can be used by states in assessing the status of waterbodies and in making listing determinations. This section, and the following three sections, primarily focus on the use of bioassessment data. Other types of biological data are discussed briefly in Section 5.3.7.

5.1.1 Using Biological Data in the Context of State Water Quality Standards

As with all assessment methodologies, the better developed the methodology, the better the tool will be in its application. Bioassessment data can be used by states to develop biological criteria, or biocriteria for their WQS. Biocriteria are numeric values or narrative descriptions that are established to protect the biological condition of the aquatic life inhabiting waterbodies of a given designated aquatic life use. Biocriteria can be formally adopted into a state's WQS and used as waterbody response criteria in a regulatory fashion similar to other water quality criteria. To date only a few states have taken this approach. More commonly, biocriteria are developed by states as quantitative endpoints to interpret their narrative biological quality standards. Most states have some form of a narrative biological condition standard formally adopted into their WQS.

Bioassessment data can also be used to establish or refine the aquatic life designated use(s) for a waterbody. By doing so, a state can develop biologically based aquatic life uses that may be more appropriately protective of the biological integrity of the waterbody than simple broad aquatic life use categories (e.g., cold water/warm water), or other uses unrelated to the natural biological quality and variability that a waterbody may be designated for. To make improvements, bioassessment data can be used to refine or tier aquatic life designated uses and to quantitatively define the level of biological condition associated with each tier. With tiered aquatic life uses, a state can set numeric biocriteria that clearly define the upper and lower bounds of biological conditions expected within each aquatic life tier. When approached in this fashion, a state will have aquatic life WQS that clearly and precisely define what the management objective is for a given waterbody and the numeric benchmarks above and below which the objective is or is not achieved.

The more biological assessment data are used to refine the aquatic life uses, to develop biocriteria to protect those uses, and to assess attainment of waterbodies against those standards, the more precise and reliable the assessments will be. Once bioassessment data are used both in designating the aquatic life use and in assessing use attainment, the more confidence a state can have in its decisions regarding waterbody status, the need for listings or the capability to de-list.establish For these reasons, states should use biological assessment data as a core indicator for revising and improving their aquatic life designated uses and criteria as long as the state can provide documentation of the adequate quality and rigor of the key elements of the state's bioassessment program.

A highly developed biological assessment program (like the Level 4 program described below), coupled with biologically based aquatic life uses and numeric biocriteria to protect those uses, constitutes the most effective combination for assessing and managing aquatic life resources, and should be the goal of all states. The more developed the bioassessment and criteria program, the more reliable and appropriate listing decisions will be, and the better and more effective the management efforts can be to restore those waterbodies.

Some states, such as Maine, Ohio, Vermont, Florida, Maryland, Kentucky, and Oregon, have already constructed biological assessment and standards programs for streams and small rivers incorporating tiered aquatic life uses derived from their bioassessment data, and are protecting those uses through numeric or narrative biocriteria. Most other states are developing programs and are at different levels of implementation and at different levels of the quality continuum described below. This guidance attempts to address all programs regardless of the stage of development, as EPA believes defensible assessment and listing decisions can be made at any stage of development or level of quality. However, the lower the level of development of the biological assessment program, the more restricted the assessment and listing decisions will be. The guidance and information described in this chapter attempts to provide recommendations for the full range of assessment and standards programs for streams and small rivers, but should also be applicable to other waterbody types as programs for those waterbodies are developed by states in the future.

Quality levels for aquatic life designated uses within WQS programs

A clear description of how biological data are used to interpret applicable WQS is an important element of a state's assessment methodology. States use a variety of approaches for integrating biological data in their WQS programs. The most common approaches are described below.

- **Level 1:** Minimal WQS program for aquatic life use protection. Possibly only one, or very few, aquatic life uses that apply to all waterbodies that are not defined or interpreted with biological assessment information. No numeric biocriteria and possibly only a generalized narrative biocriterion without defined implementation procedures or translator mechanisms. The majority of the key bioassessment elements (see Section 5.2) are also Level 1.

- **Level 2:** Basic WQS program for aquatic life use protection. A state has aquatic life designated uses with different categories or subcategories related to recreational fisheries, cold water/warm water fisheries, species of concern, or other descriptions. Biological assessment approaches have not yet been developed that define attainment of the aquatic life uses. No tiering of aquatic life uses using the bioassessment data. A biocriteria index has been developed for interpreting the bioassessment data, and the index is used to interpret the state's narrative biocriterion that applies to all waterbodies and all uses, but the index has not been related to the designated aquatic life uses. Most of the key bioassessment elements are at Level 2 or above.

- **Level 3:** Intermediate WQS program for protecting aquatic life use. A state, territory, or authorized tribe has developed bioassessment protocols and has derived a biocriteria index

for interpreting the bioassessment data and implementing their narrative criteria. Numeric biocriteria, however, are not yet adopted, but the state has adopted well-defined biologically based designated uses for their waterbodies and also has specific biological descriptions or methods used to define the uses. Rather than applying one narrative biological criteria to all uses, the state may have tiered narrative biocriteria for each use that are measured by bioassessment data. The tiered narrative biocriteria are adopted into their WQS, and well-described implementation procedures or translator mechanisms that define quantitative thresholds are described either in the WQS or in other implementing regulations or policies and procedures documents such as the continuing planning process or consolidated assessment and listing methodology.. Most of the bioassessment key elements are at Level 3 or above.

- **Level 4:** Advanced WQS program for protecting aquatic life use. A state, territory, or authorized tribe has tiered aquatic life uses developed using bioassessment data that reflect a continuum of biological conditions based on regional reference conditions and natural biological integrity. Numeric biocriteria are adopted in the WQS for each tiered aquatic life use, and well-described implementation procedures or translator mechanisms that define quantitative thresholds are included either in the WQS or in other implementing regulations or policies and procedures documents such as the state, territory, or authorized tribe's continuous planning process or consolidated assessment and listing methodology.. A Level 4 program includes a monitoring and assessment program that accurately and precisely assesses the quality of biological conditions in any given waterbody and compares it against the aquatic life use tiers and the biocriteria thresholds. Most of the key bioassessment program elements are at Level 4 or better.

EPA recommends that states use biological assessments to refine, or tier, their aquatic life uses. A tiered approach to classification should articulate appropriate ecological expectations for state waters (e.g., reference conditions) and specify goals for individual waterbodies (e.g., tiered, designated aquatic life uses). Appropriate water quality criteria may then be adopted into state standards to protect the specific designated uses. The water quality criteria and any needed implementation procedures should provide for quantifiable measurement of each specified use. This approach will better protect high-quality waters, provide for more accurate evaluation of effectiveness of controls and best management practices, and enhance public confidence and participation in the WQS-setting and waterbody listing process.

The states of Maine, Vermont, and Ohio have well-described use classification systems in their standards. Currently, most states are using or preparing to use Level 2 or 3 (i.e., they have adopted narrative biocriteria and either have well-developed bioassessment procedures in place or are in the process of validating procedures and decision thresholds). Many states have rigorous bioassessment programs that can serve as a basis for implementing or adopting numeric biocriteria in their water quality standards. As of 1996, all but three states had either developed, or were in the process of developing bioassessment approaches for streams. Thirty states used bioassessment to interpret aquatic life use attainment and 28 states had narrative biocriteria (U.S. EPA 1996a). Today only a few states have numeric biological criteria in their standards (e.g., Maine, Ohio, and Florida for streams; Delaware for estuaries). EPA is updating the status of

state and tribal bioassessment programs. Preliminary indications are that state and tribal program growth and sophistication have continued beyond 1996 levels.

5.1.2 Using Bioassessment Data To Determine Impairments in a Level 1 Program

For states with a Level 1 aquatic life standards program and a Level 1 bioassessment program, using bioassessment data to determine and list impairments may be tenuous, with relatively low precision. Level 1 bioassessments can detect severe impairments but have less power to distinguish degrees of impairment or degrees of biological recovery, and therefore tend to provide all-or-nothing results. Because of this, assessments and listing based on Level 1 programs require a significant amount of best professional judgment and scientific interpretation. Highly altered waterbodies may be reasonably listed as impaired, but determining recovery and restoration of the waterbodies, without reference conditions or tiered biological quality thresholds derived from higher level bioassessment data, may be more difficult. Level 1 bioassessment data may best be used to screen waterbodies for further study (i.e., those waterbodies with severe impairments revealed using Level 1 bioassessments should be listed as priority waterbodies for more intensive bioassessments, conducted at Level 2 or above, before actually being listed). In addition, Level 1 bioassessment programs generate data that generally should not be used to conclude a waterbody is attaining WQS, where other aquatic life data show exceedances of 304(a) aquatic life criteria, etc.

5.1.3 Using Bioassessment Data To Determine Impairments in a Level 2 Program

In a state with a Level 2 aquatic life use protection program, assessment and listing decisions using bioassessment data may involve some scientific and regulatory judgment using thresholds for attainment that are derived from the bioassessment data as interpretations of the narrative aquatic life standard. For a state with a Level 2 aquatic life WQS program and having only one general aquatic life use for all or most of its waterbodies, and also having a Level 2 or higher bioassessment program, decisions should be made as to where along a continuum of biological condition the state would identify a threshold level that is considered acceptable, and therefore in attainment of the standard. A quartile approach may be useful for determining attainment thresholds from the bioassessment data (see Section 5.3).

Under such an approach, once quantitative thresholds are established from the bioassessment data, impairment occurs when a bioassessment of a waterbody shows a statistically significant departure of biological condition from the threshold. (Oregon, Kentucky, and North Carolina now use this approach.)

For a state with a Level 2 aquatic life standards program, it may be necessary to document the procedures and rationale for interpreting the narrative standard and the statistical derivation of the decision thresholds that were derived from the bioassessment data. For additional guidance on this, see Section 5.3.

5.1.4 Using Bioassessment Data To Determine Impairments in a Level 3 Program

When a state has not yet adopted numeric biocriteria, but has adopted well-defined biologically based designated uses for its waterbodies with specific biological descriptions or methods that define the biologically based uses, a state may conclude that impairment occurs when a biological assessment shows that the waterbody is not achieving the biologically based designated use in accordance with the state's methods and procedures. Usually these states will have a well-defined quantitative threshold in their regulations or are implementing procedures that define the upper and lower bounds of biological condition acceptable for each aquatic life use tier derived from their bioassessment data. (Vermont and Maine currently use this approach.)

5.1.5 Using Bioassessment Data To Determine Impairments in a Level 4 Program

If a state has adopted numeric biocriteria in its WQS for all waterbodies that numerically define the range of biological condition for each of their aquatic life use tiers, the state may conclude that impairment occurs when the biological condition of a waterbody with a given designated aquatic life use is significantly less than the numeric biocriterion defining the lower end biological condition threshold for that use. With a Level 4 aquatic life standards program, determining waterbody impairments is more definitive because of the quantitative precision the numeric biological thresholds provide. As such, a state should have a high level of confidence in the listing decisions it makes. Likewise, adequate restoration of a waterbody, as a result of a TMDL or other management action, is readily determined when bioassessments in the restored waterbody show that the biological condition has improved to above the lower numeric biological threshold for the particular aquatic life use. In this case, delistings of restored waterbodies should become a straightforward process. With tiered aquatic life uses and numeric biocriteria thresholds defining the highest and lowest acceptable biological conditions within the aquatic life use tier, it is also feasible to track biological degradation in waterbodies. As a waterbody's biological condition is tracked and the waterbody begins to exhibit a condition that is approaching the lower level for the aquatic life use tier, the waterbody may be listed as "threatened" so that necessary actions can be taken to prevent the waterbody from deteriorating any further. (Ohio is one state using this approach.)

5.2 How does the State Use the Key Elements of a State Biological Assessment Approach to Assess and Document Data Quality, Including the Use of Other Data?

The rigor and quality of the biological data are integral to a biological assessment program. Depending on how they are derived, not all biological data are necessarily of equal value for assessing WQS attainment/impairment. The following sections outline the key elements that should be included in a state bioassessment program to ensure that good-quality data are the basis from which reliable attainment/impairment decisions are made as well as decisions regarding sampling and monitoring design (5.2.1), classification of waterbodies (5.2.2), choice of reference conditions (5.2.3), choice of indicator assemblages (5.2.4), choice of field and laboratory protocols (5.2.5), and precision of the biological methods (5..6). Table 5-1 provides a summary of the key elements for the four levels of rigor in conducting bioassessments.

Table 5-1. Defining levels of rigor for key components of a biological assessment

	Level 1 information	Level 2 information	Level 3 information	Level 4 information
Temporal Coverage	No index period is identified and sampling can be scattered throughout the year. This approach is not recommended because it does not help to establish a reliable benchmark reflecting the natural cycles of spawning, recruitment, migration, and mortality.	A seasonal period is identified for convenience in sampling or to match existing programs. Sampling outside the index period may be done, but usually is reserved for emergency response monitoring.	A well-documented seasonal index period(s) is identified or coverage is comprehensive (periodic sampling occurs throughout the year). Index periods are selected based on known ecology to minimize natural variability, maximize gear efficiency, and maximize the information gained on the assemblage (U.S. EPA 1999a). Reference conditions are calibrated for the index period(s).	A well-documented seasonal index period(s) identified; multiple sampling during index period likely; reference conditions calibrated for the index periods.
Natural Classification	No classification of ecosystems. This approach is not recommended, because natural variability is not partitioned to improve the benchmarks for assessment.	Minimal classification limited to individual watersheds or basins. This approach may not recognize stream continuum principles where headwaters differ in function from mainstem. In estuaries and lakes, classification may apply only to portions or embayments.	Classification recognizes geographical or other similar organization. This approach usually is based on landscape features and supplemented with instream or other waterbody characteristics.	Classification based on a combination of landscape features and physical habitat structure of waterbody type. This approach provides the best classification scheme for assessment.
Reference Conditions	No reference condition formally established. Presence and absence of key taxa may constitute the basis for assessment. Professional opinion may be used to support assessment of attainment. This approach may be difficult to defend, especially in listing determinations, than those relying on more formal scientific evidence.	Reference conditions preestablished by professional biologist and based on known ecology of area. A site-specific control or paired watershed approach may be selected for assessment. Regional sites generally are not used at this level.	Reference conditions may be site-specific, but normally are based on watershed-scale assessments. Regional reference sites have likely been developed for the relevant waterbody type and are the basis for assessment and monitoring.	Regional reference conditions are established for each waterbody class and consist of sites and/or other specified means of establishing regional expectations for assessing and monitoring each waterbody.

Table 5-1. Defining levels of rigor for key components of a biological assessment (continued)

	Level 1 information	Level 2 information	Level 3 information	Level 4 information
Indicator Assemblages	Visual observation of biota; poor taxonomic resolution.	One assemblage (usually of an invertebrate); adequate but consistent taxonomic resolution.	Single assemblage collected and analyzed; high data quality and higher taxonomic resolution.	Two or more assemblages collected and analyzed; taxonomic resolution to the lowest practical taxon (mostly genus/species).
Field and Lab Protocols	Documentation of methods is cursory, and the methods usually are not written as SOPs. Methods may be highly variable, relying primarily on best professional judgment.	Methods generally are well documented, but QA/QC may be minimal. Training of biologists may be oriented to new or inexperienced staff only.	Methods are well documented and SOPs are updated periodically. An effective QA/QC program is in place. Training is provided periodically throughout the year for all staff to raise skill levels and enhance interaction and consistency.	Same as level 3, but methods cover multiple assemblages.
Precision of Assessments	Precision of method is low or not measured. Replicate data for estimating precision are not normally available. Capability of indicator to distinguish between human and natural influences is unknown.	Precision of method is moderate. Method is better documented to enable more consistent sampling and higher precision. Capability of indicator to distinguish between human and natural influences has been determined based on studies conducted in other states or regions.	Precision is moderately high, maintained through rigorous methods, training, and periodic refinements or improvements to the implementation of the methods. Capability of indicator to distinguish between human and natural influences has been documented within the state, but is generally based on impaired and reference sites without a gradient of stressors/human influence.	Normally precision is highest, reflective of the most rigorous methods development and QA/QC, with good repeatability in assessments and a high level of confidence in analytical results. Capability of indicator(s) to distinguish between human and natural influences is quite high and based on a gradient of stressors/human influence, which may also include impaired and reference sites.

Table 5-1. Defining levels of rigor for key components of a biological assessment (continued)

	Level 1 information	Level 2 information	Level 3 information	Level 4 information
Thresholds	No formal index or community-based endpoint. Assessment may be based only on presence or absence of targeted or key species. (Some citizen monitoring groups use this level.) Attainment thresholds not specified. .	A biological index or endpoint is established for specific waterbodies, but is likely not calibrated to waterbody classes or statewide application. Index is probably relevant only to a single assemblage. Watershed monitoring can be used where regional reference conditions have not been established. Attainment thresholds may be based on dividing the total possible index or model score into equal parts (e.g., quarters, thirds).	A biological index, or model, has been developed and calibrated for use throughout the state or region for the various classes of a given waterbody type. The index is probably relevant only to a single assemblage, but may or may not be applicable to several states or tribes. Several states conduct assessments using Level 3 information (e.g., Florida, Arizona). Attainment thresholds in these states are based on discriminant model or distribution of candidate reference sites.	Biological index(es), or model(s) for multiassemblages is (are) developed and calibrated for use throughout the state or region. Integrated assessments using the multiple assemblages are possible, thus improving both the assessment and diagnostic aspects of the process. (Ohio and Idaho are examples of states using this approach.) Attainment thresholds in these states are the same as Level 3, except power analysis is used to determine the number of assessment categories.

5.2.1 Index Period or Other Temporal Conditions During Which a State Collects Biological Data

As part of its monitoring program design, a state should clearly identify waterbodies of interest by including them in what is called a target population. In the biological assessment program, such identification typically is done by waterbody and ecoregion type, along with selection of an index period. Because it may not be possible to adequately monitor each waterbody or waterbody type, most monitoring programs collect data from a representative sample of waterbodies in a target population (e.g., EMAP, MBSS). If the monitoring program takes a well-designed sample survey approach or a comprehensive, nonrandom approach (as Ohio does), the state may obtain statistically valid inferences about the condition of the target population.

A state should also document the index period (time of year and duration) when it will sample the condition of the biological community, or specify that it will sample year-round. EPA recommends establishing index periods for a particular season, time of the day, or other window of opportunity when signals are determined to be strong and reliable. Further, EPA recommends that only results from similar index periods be compared when assessing WQS attainment/impairment.

The sampling does not need to occur during the more severe or worst-case conditions. However, understanding the dynamics of how an ecosystem functions at different times enables an investigator to better interpret data from prescribed index periods. The use of an index period also allows a better concentration of sampling during a period when reference conditions have been characterized. A specified index period is used in most state bioassessment programs, although the level of specificity varies.

5.2.2 Natural Classification of Waterbodies

The state should clearly document how it determines the natural variability of its biological data. Classification is useful in evaluating natural variability and distinguishing it from variability resulting from human-induced changes. Classification of waterbodies may be based on waterbody type (e.g., rivers, streams, lakes, wetlands, estuaries), watershed drainage size, ecological regions, elevation, temperature, and other physical features of the landscape and/or waterbody. The number of classifications the state can analyze may be limited by the number of samples taken and the availability of candidate reference sites within each class. EPA recommends classifying more specifically than simply by waterbody type, because it is highly unlikely that the biological condition of any given waterbody type is uniform throughout the entire state. States should list the classification approach(es) used, if any, for all waterbody types monitored.

Ecoregions have been used successfully as primary classification schemes (e.g., in Ohio; see Yoder and Rankin 1995), or as aggregates of ecoregions (for example, in Florida, see Barbour et al. 1996; in Wyoming, see Gerritsen et al. 2000). Ecoregions are areas of relative ecosystem homogeneity (or similar quality) defined by similarity of land form, soil, vegetation, hydrology, and general land use. For example, streams of a given ecoregion are more similar to one another

than they are to streams in another ecoregion. In coastal marine areas, large-scale biogeographic provinces are similar in concept to ecoregions. These provinces are characterized based on latitude, climate, and similarities in land form (Holland 1990). For wetlands, classifying by hydrogeomorphic type has been used by many states in evaluating natural variability among wetland types. For a discussion of various methods of classifying wetlands see *Methods for evaluating wetland condition: wetlands classification* (U.S. EPA 2002a).

Ecoregions are not the only classifications of freshwater ecosystems; Hawkins et al. (2000a) point out that the amount of biotic variation related to landscape features is not large, and augmenting classifications based on local habitat features accounts for substantially more variation than the larger-scale environmental features. Some states have used other landscape factors such as elevation and rainfall to classify their waterbodies (Spindler 1996).

5.2.3 Reference Conditions

Reference conditions should be defined to assess a waterbody's ecological health and establish water quality goals. Reference conditions serve as the benchmark of biological integrity against which a waterbody's conditions are compared. The assessment and listing methodology should describe how the state developed its reference conditions and whether they are based on assessment of reference sites or were developed through other means. The assessment and listing methodology may incorporate by reference the state's biological assessment methods and indicate which of the four levels of rigor best characterizes those methods.

State assessment methodologies should clearly document how reference sites are selected and used. A reference condition can be derived from reference sites, an empirical model of expectations that may include knowledge of historical conditions, or a model extrapolated from ecological principles. Normally, actual sites that represent best attainable conditions of a waterbody are used. Generally, EPA recommends the use of a regional reference condition based on an aggregate of sites that allows for broader application in state water resource programs than individual, site-specific conditions (U.S. EPA 1996a).

Where reference sites are not available (e.g., for large ecosystems such as rivers, estuaries, near-shore coastal areas, and in significantly altered systems such as urban centers and cropland areas), a disturbance gradient might be constructed to extrapolate to an appropriate reference condition (Karr and Chu 1999). This approach requires some knowledge of both stressor gradients and biological condition gradients.

Abiotic factors also may be used in selecting candidate reference sites. Use of these factors helps to avoid circularity in defining biological characteristics that become the basis of reference conditions. Candidate reference sites then are evaluated to determine the degree of human modification that has occurred. Factors considered may include human population density and distribution, road density, and the presence of mining, logging, agriculture, urbanization, grazing, or other land uses. This information can be gleaned from GIS data layers, maps, and/or evaluations by resource managers. Candidate sites should be eliminated if they have undergone extensive human modification, especially to riparian zones. Candidate sites can be selected by

from probabilistic sampling (*a posteriori* determination) or from targeted sites (*a priori* selection).

Abiotic selection criteria can range from a few chemical criteria to a whole range of factors as discussed above. The rigor of the criteria also varies from very conservative, which may restrict the number of candidate reference sites selected, to very liberal, which may increase the number. Although EPA prefers a conservative approach, states may take different approaches based on their knowledge of the reference sites. EPA suggests using a conservative approach when greater uncertainty exists as to whether the candidate sites are likely to represent the highest quality waters. State methodologies should include documentation of these decisions.

It is very important that the state or tribe verify in the field the current conditions of candidate reference sites. A candidate site should be eliminated if conditions preclude its ability to serve as a reference for high-quality water. A reference site may be natural, minimally impaired (somewhat natural), or best available (altered system).

In summary, when reference sites are used to establish reference conditions, the state should document how (by what criteria) it selects reference sites and how it uses them to define regional reference conditions (e.g., by combining sites in a regional reference condition, or through other approaches such as a paired watershed or upstream/downstream design).

5.2.4 Indicator Assemblages

State assessment and listing methodologies should document both the assemblages used as indicators and the level of taxonomy used to assess them. Biological indicators can be separated into four principal assemblages that are used for assessing WQS attainment/impairment decisions: benthic macroinvertebrates, fish, algae, and aquatic macrophytes. Research is under way on birds and amphibians as candidate assemblages for wetlands, marshes, and headwater and ephemeral streams, as well as other waterbody types (U.S. EPA 2001- MAIA).

Although a single assemblage may be sufficient to make a WQS attainment determination, EPA recommends the use of more than one to enhance confidence in the assessment finding. Each assemblage serves a different function in the aquatic community, has differing habitat ranges and preferences, and may be susceptible to stress in varying manners and degrees. Several states routinely collect and analyze more than one assemblage in their water quality assessments, although different agencies within a state may collect the data.

Benthic macroinvertebrates

The benthic macroinvertebrate assemblage inhabits the sediment or bottom substrates of waterbodies and responds to a wide array of stressors in different ways. Often it is possible to determine the type of stress that has affected a macroinvertebrate community (U.S. EPA 1990a, 1999a). Because many macroinvertebrates have life cycles of a year or more and are relatively immobile, macroinvertebrate community structure generally is a function of past conditions in

the specific waterbody. The benthic assemblage is the most common assemblage used in bioassessments for state water quality programs (U.S. EPA 1996b).

Taxonomy: Genus/species taxonomic identification provides the most representative information on ecological relationships and best resolution in sensitivity to impairment (U.S. EPA 1999a). In the Northwest, it is standard practice in bioassessments for all macroinvertebrates in the subsample to be identified to the lowest possible taxonomic level, generally genus or species (Hayslip 1993). However, in some geographical regions of the United States, family-level identification is used more commonly and may be sufficient for assessments (Hawkins and Norris 2000, Bailey et al. 2001). The scientific determination of level of taxonomy should include a knowledge of adaptive radiation within the fauna (i.e., estimates of the number of genera and/or species per family). For example, the higher the ratio of genera to families, the less likely a family-level identification approach will be adequate. Naturally depauperate systems, such as low-gradient streams or oligotrophic lakes, may warrant family-level indices. In lakes and estuaries, biomass measurements are done on taxonomic groupings (e.g., family or genus) as part of bioassessments.

Whatever level of taxonomic rigor is chosen, the state should clearly document it in its assessment and listing methodology. A macroinvertebrate "voucher collection" for each major basin, ecoregion, site class, or other appropriate study unit is recommended highly. Such a collection contains a representative of each taxon and serves as a basin record and reference for checking identifications as well as a providing a data quality check. A senior aquatic taxonomist should check the specimens entered into the collection for accurate identification and, if necessary, send them out to recognized experts for verification. Ideally, the voucher collection should be housed in a museum or university. The state's protocols for establishing and maintaining such a collection also need to be described (or referenced) in its assessment and listing methodology.

Fish

A bioassessment conducted using a fish assemblage requires that all fish species (and size classes), not just game fish, be collected. Fish are good indicators of long-term effects and broad habitat conditions because they are relatively long-lived and mobile (Karr et al. 1986). The fish assemblage also integrates various features of environmental quality, such as food and habitat availability. The physical degradation of streams can cause changes in the food web and the composition and distribution of habitats (Lonzarich 1994). The objective of the fish assemblage portion of any protocol is to collect a representative sample using methods designed to (a) collect all except rare species in the assemblage and (b) provide a measure of the relative abundance of species in the assemblage. Fish assemblages in streams are used more commonly in bioassessments conducted in the eastern and midwestern United States than elsewhere, although some programs in other regions are investigating their utility (U.S. EPA 1996b). Fish assemblages in the streams and rivers of the western United States have been the subject of fewer studies because of their more depauperate nature. Also, fish diversity is low naturally in headwaters and other small streams, as well as in intermittent streams, making fish less viable indicators than other assemblages. Fish are considered important indicators in larger waterbody

types (e.g., lakes, estuaries); however, here too, fish assemblages have been used less often than other methods in water quality assessments because of mobility and sampling difficulties in these waterbodies (U.S. EPA 1998, 2000a).

Taxonomy: All fish species should be identified to species level either in the field or the laboratory, depending upon the expertise of the field crew. As with benthic macroinvertebrates, it is important to retain voucher specimens (ideally in a museum or university), and EPA recommends that a taxonomic expert verify and make determinations on any problematic taxa. Additional information on species of interest may be obtained by recording total length and weight. In addition, fish may be examined for external anomalies.

Periphyton or phytoplankton

Algae are primary producers and responsive indicators of environmental change. The periphyton assemblage serves as a good biological indicator in streams and shallow areas because of its naturally high number of species and rapid response to exposure and recovery. Most algal taxa can be identified to species level by experienced biologists, and the tolerance or sensitivity to specific changes in environmental conditions is known for many species (Rott 1991, Dixit et al. 1992). Because periphyton is attached to the substrate, this assemblage integrates physical and chemical disturbances to a stream reach. However, few state environmental agencies have developed protocols for the periphyton assemblage in streams. Recently, Idaho proposed a method to use diatoms in assessing the biointegrity of large Idaho rivers (IDEQ 1999). Phytoplankton is a common assemblage used in lake (U.S. EPA 1998) and estuary (U.S. EPA 2000a) assessments**.**

Taxonomy: In general, EPA recommends identifying algae to the species level in rivers and wadeable streams because (1) differences among assemblages that may occur at the species level will be better characterized and (2) large differences exist in ecological preferences among algal species within the same genus. However, substantial information can be gained just by identifying algae to the genus level. Although valuable ecological information may be lost, the costs of genus-level analyses are less, especially if inexperienced analysts are involved (U.S. EPA 1999a, Chapter 6).

Identifying diatom genera in assemblages can provide valuable characterizations of biotic integrity and environmental conditions, and may be a good approach when implementing a new program and only an inexperienced analyst is available. As the analyst gains more experience counting, the taxonomic level of the analyses should improve. Eventually, the cost of counting and identifying algae to the species level becomes not much greater than the cost of analysis to the genus level (U.S. EPA 1999a, Chapter 6).

For assessing lakes, EPA recommends sampling the phytoplankton assemblage and counting and identifying cells to the order or genus level. Simplified field and laboratory procedures are possible for measurements based on higher taxonomic levels such as division or order. Identification to the species level is considered supplemental at this time, because it is not clear

that the information gained represents a substantial improvement over higher levels of taxonomy (U.S. EPA 1998).

Aquatic macrophytes

Aquatic macrophytes include vascular plants (grasses and forbs) and may be emergent or submergent. Vascular aquatic macrophytes are a vital resource because of their value as extensive primary producers and habitat for fish and waterfowl (U.S. EPA 2000a). As an ecological indicator, this assemblage is most important in estuaries (U.S. EPA 2000a) and wetlands (U.S. EPA 2002). Excessive nutrient loadings lead to prolific phytoplankton and epiphytic macroalgal growth on grasses that outcompete the macrophytes (U.S. EPA 2000a).

Taxonomy: Macrophytes are identified to species level or categorized as emergent, submergent, or floating leaf for purposes of assessment. The taxonomy serves a basis for areal coverage or standing crop biomass analyses. Because submerged aquatic vegetation distributions are specific to a given habitat parameter (i.e. salinity, depth, etc.), they most commonly occur in monotypic stands with some mixed beds (e.g., *Zostera and Ruppia*). In these cases the taxonomy analysis will be less revealing than using abiotic parameters as early warning measures (e.g., light attenuation coefficient, total suspended solids, chorophyll *a*).

Whatever assemblage(s) are used, states should document the rationale for choosing them and include in their assessment and listing methodologies the value and purpose of the assemblage(s) in attainment and listing decisions. If not documented elsewhere, the consolidated assessment and listing method is the appropriate place to document these decisions. The scientific credibility of every assessment depends on how the assemblage is selected.

5.2.5 Field and Laboratory Protocols for Indicator Assemblages

Standardization of laboratory and field methods should be done to establish the validity and reliability of biological data. Whatever assemblage is chosen, the methods for sample collection and laboratory analysis should be documented fully. EPA has published a generic quality assurance (QA) project plan guide for programs using community-level biological assessment in wadeable streams and rivers (see U.S. EPA 1995). The development of standard operating procedures (SOPs) for field and laboratory methods should include an effective quality assurance program with quality control (QC) checks. To minimize bias, reduce error, and maintain a high level of data integrity, the SOPs and QA/QC plan should identify the specific procedures for all aspects of the biological program.

Information on data quality objectives and quality assurance/quality control procedures usually is documented in a separate quality assurance project plan and standard operating procedures document, which can be referenced in the state's general assessment and listing methodology. This information should be available for other parties to use as a reference in developing compatible monitoring projects.

Considerations for macroinvertebrate assemblage sampling and laboratory analysis

Habitat type: The three basic macroinvertebrate habitat types available to sample are (a) artificial substrate, (b) multihabitat, and (c) single habitat. Each type offers sampling advantages and disadvantages. Some choices are more appropriate in some regions of the country than in others. State assessment methodologies should describe which habitat they are sampling and why they have chosen it.

Each of the three habitat types is commonly used throughout the United States to sample aquatic organisms. However, at a minimum the following considerations should be met when selecting which one to sample: (1) adherence to strict quality control procedures to provide consistency and avoid sampling error, (2) reliance in choosing a single habitat type based on its availability and dominance as a productive organism habitat (e.g., cobble in streams, kelp beds in coastal areas, or mud in estuaries), (3) preference for a multihabitat approach in systems with diverse habitat, and (4) use of artificial substrates, which leads to sampling habitat that is natural for the system(s) under study (e.g., rock baskets in cobble streams or lakes, or multiplate Hester Dendy substrates to represent woody debris in streams). A state's assessment methodology should describe which habitat type it is sampling and why it was chosen.

Gear/number of samples: In streams, macroinvertebrate samples usually are taken with either a Surber sampler, Hess sampler, D frame net, or artificial sampler. State methodologies need to specify the gear type to be used. In addition, they need to document the specific characteristics of that gear (e.g., the standard mesh size for nets, if applicable) and the number of samples taken from the habitat type.

For riffle sampling, EPA's Rapid Bioassessment Protocols (RBPs) recommend sampling a minimum of 2 or 3 m^2 of stream bottom (U.S. EPA 1999a). The RBPs recommend compositing (combining) riffle samples into a single sample representative of the stream reach; however, replicates are taken at a proportion of the sites (usually 10% of the sites) to measure sampling precision (U.S. EPA 1999a). Others (Kerans et al. 1992) recommend taking replicate samples at all sites (i.e., taking more than one sample from a stream reach and keeping it separate for taxonomic identification and enumeration). Three to five replicates are commonly used at each site in many research studies (Resh and McElvary 1993), though scientific debate continues on the appropriate number of samples per site/reach. The same approach (i.e., compositing samples with replicates for precision estimates) is recommended for lakes (U.S. EPA 1998) and estuaries (U.S. EPA 2000a) (however, the gear for infaunal sampling consists of grab samplers (e.g., Ponar). Again, state assessment methodologies should document (or reference) their sampling approach.

Subsampling: Bioassessment programs designed to support assessing WQS attainment/impairment decisions rely on timely and cost-effective laboratory processing of benthos samples. Alternatively, analysts sometimes use a predetermined fraction of the field sample for identification and enumeration, called "subsampling," the goal of which is to provide an unbiased representation of a larger sample (Barbour and Gerritsen 1996). Crucial to the reduction of costs and time associated with processing benthic samples, the subsampling

procedures developed by Hilsenhoff (1987) and modified by Plafkin et al. (U.S. EPA 1989) have been implemented in many state programs. As an improvement to the mechanics of the technique, Caton (1991) designed a sorting tray and method that allow for rapid isolation of organisms and easy removal of all organisms and debris as well as the elimination of any investigator bias in the process. In Rocky Mountain streams of Wyoming, a 200-organism subsample was found to be optimal in terms of information return for the investment (Gerritsen et al. 1996). Most agencies in the Northwest use either a 300- or 500-organism subsample. However, proportional subsampling may be a viable alternative to fixed-count subsampling, and has been advocated as more accurate in some cases (Courtemanch 1996, Cuffney et al. 2000).

Whatever procedure and number of organisms are subsampled for identification, the state's assessment methodology should clearly document (or reference) the approach used. Precision estimates are important to help interpret results from subsampling efforts. An approach whose precision is considered low indicates lower confidence in the interpretation of data than one whose precision is considered high. For instance, subsampling 100 organisms, as opposed to 300 or 500, will provide less information about taxa richness because the probability of capture is less. However, knowing with precision how taxa richness is estimated from only 100 organisms may, in limited circumstances, still allow an agency to adequately assess the condition of a site. EPA recommends that states test the level of subsampling and establish precision measurements for application to their subsampling levels.

Considerations for fish assemblage sampling and laboratory analysis

Reach length or sampling area: The most recent revision to the Rapid Bioassessment Protocols (U.S. EPA 1999a) describes two acceptable methods for site or reach selection. The first is a fixed distance method such as that used by Ohio EPA (150–200 meters) and Massachusetts DEP (100 meters). The second is a proportional distance method such as that used by the EPA Office of Research and Development's EMAP program (40 times the stream width). In lakes and estuaries, fish sampling is to occur in the littoral zone along the shoreline, or in the pelagic areas for a specified distance or time (U.S. EPA 1998, 2000a).

Field methods: The RBPs recommend electrofishing as a standard sampling technique for use in streams and small areas (U.S. EPA 1993a). Single-pass removal through electrofishing is sufficient to obtain a representation for biological assessments (Bauer and Burton 1993). However, in some cases electrofishing may not be allowed in order to accommodate the presence of endangered species, or may not be practical for other reasons. In these cases, other methods such as snorkeling or seines are used. Snorkeling may miss some smaller, nongame species of fish and therefore is less useful for assemblage-level analysis. Sampling gear used in large waterbodies, such as rivers, lakes, and estuaries, consists of seines, gill nets, and trawls. The method selected should be documented in the assessment methodology.

Considerations for periphyton and phytoplankton assemblage sampling and laboratory analysis

Field methods: The two major categories for periphyton sampling differ as to the type of substrate sampled (natural versus artificial). For an accurate assessment of the assemblage, samples should be collected during periods of stable instream flow.

For natural substrates, samples may be collected from either all available microhabitat types or from a single habitat type. The procedures for sampling from all available microhabitats have been adapted from the Kentucky and Montana protocols (Kentucky DEP 1993, Bahls 1993) and are reported in the latest version of the RBPs. An alternative to compositing several microhabitats is to select a single habitat type that sufficiently characterizes the study reach. The most accurate way to decrease sample variability is to collect from only one type of habitat within a reach and to composite many samples within that habitat (Rosen 1995). If multiple habitats are sampled, the samples should be kept separate, by habitat, for analysis.

Periphyton also can be sampled by collecting from artificial substrates that are placed in aquatic habitats and colonized over a period of time. This procedure is especially useful in larger (nonwadeable) streams, rivers with no riffle areas, wetlands, and lake environments. Kentucky (Kentucky DEP 1993), Florida (Florida DEP 1996), and Oklahoma (Oklahoma CC 1993) have used this technique successfully. Either surface (floating) or benthic (bottom) periphytometers are used and fitted with glass slides, glass rods, clay tiles, plexiglass plates, or similar substrates that occur in the study area. The minimum requirements for periphyton investigations are as described in the Rapid Bioassessment Protocols (U.S. EPA 1999a) for streams. The minimum requirements for phytoplankton investigations are as described in the Lakes Bioassessment and Biocriteria Document (U.S. EPA 1998) and the Estuarine Bioassessment and Biocriteria Document (U.S., EPA 2000a).

Phytoplankton standing stock is estimated by chlorophyll *a* measurements. One approach is to collect three replicate samples at each station at one-half the Secchi depth using a Kemmerer or Van Dorn sampler (U.S. EPA 2000a). Another approach is to collect a depth-integrated sample through the entire photic portion of the water column. The same techniques for phytoplankton collections are applicable to lakes and reservoirs (U.S. EPA 1998) and estuaries and coastal marine waters (U.S. EPA 2000a).

Laboratory analysis: Generally, two types of algae can be identified for assessment: soft algae (nondiatoms) and diatoms. Some states identify the diatoms only. For data on diatom abundance, EPA recommends counting a minimum of 300 to 500 valves or frustules and recording taxa and number counted on bench sheets. Chlorophyll *a* also is analyzed in conjunction with taxonomic identification. Chlorophyll *a* is analyzed fluorometrically or spectrophotometrically following disruption of cells (by grinding) and extraction with acetone (APHA 1992). Once again, documentation of the methods selected by the state and adequate QA/QC procedures to ensure that high quality data are available for making WQS attainment decisions are important.

Considerations for macrophyte assemblage sampling and laboratory analyses

Field methods: For large waterbodies (i.e., large rivers, lakes or reservoirs, wetlands, estuaries or coastal marine areas), areal coverage and distribution of submerged aquatic macrophytes may be estimated from aerial photographs, if available, and ground-truthed at the site (U.S. EPA 2000a). The dominant taxa may be field-identified from vegetation samples collected in shallow waters. Detailed macrophyte monitoring and assessment procedures are included in U.S. EPA (1992), Ferguson and Wood (1994), and Orth et al. (1993). Macrophyte surveys in streams and wetlands usually require site visits to identify the diversity of species and delineate the areal coverage and standing crop biomass.

Laboratory analysis: Most identifications of macrophytes are done in the field. However, voucher collections and samples for biomass determinations are returned to the laboratory.

5.2.6 Precision of Biological Methods

State methodologies should document the capability of selected biological indicators to distinguish between human and natural influences. The value of a biological index lies in its capacity to be used reliably as a signal of environmental degradation. The capability of the indicator to discern differences among sites along a known gradient of disturbance should be examined critically.

The discriminatory capability of the indicator or index is determined by observing its response to environmental stress. The preferred way to do this testing is by establishing a gradient of stress based on nonbiological factors such as contaminant concentrations, physical habitat quality, or land uses (Karr and Chu 1999). Alternatively, binomial discriminatory capability can be determined by comparing biological differences between high-quality reference sites and stressed sites (U.S. EPA 1999b). Engle (2000) and McCormick and Peck (2000) address discriminatory capability for estuarine and freshwater systems, respectively. The document Evaluation Guidelines for Ecological Indicators (U.S. EPA 2000b) and the revised Rapid Bioassessment Protocols (U.S. EPA 1999a) also address this issue.

Whatever assemblage or combination of assemblages is used, the state's assessment and listing methodology should document its value and purpose in making WQS attainment and listing decisions. Fundamental requirements for a biological assessment include understanding the performance of the method (e.g., bias and precision) as well as the effects of natural variability on the method's ability to detect a gradient of environmental impairment. Biological assessments are most useful when the sample is representative of the site examined and the assemblage measured, the data are an accurate reflection of that sample, and the methods distinguish natural and measurement variability (i.e., "noise") from a true environmental effect (i.e., "signal").

Method precision indicates the level of confidence in a site characterization, partly through the liklihood that the assessment could be replicated. Precision in a bioassessment requires consideration of variability resulting from both human and natural sources. Therefore, each step

in the sampling and analysis process, including sampling precision, laboratory sorting precision, and taxonomic identification precision (ITFM 1995), should be addressed.

Bias also is an important consideration. Certain sampling gear or procedures, for example, are biased in terms of the types of biota they collect or the types of environmental conditions in which they are most efficient. It is important to understand such sources of bias and how they may interact with natural sources of variation (e.g., flow, season, geomorphology) to influence site characterization. Quality assurance programs encourage the continued documentation of variability to ensure the ability to detect long-term trends. An ongoing quality assurance program also is useful for periodically reevaluating the performance of the indicator and the adequacy of reference conditions.

Two fundamental characteristics of a biological assessment are that samples are representative of the site or assemblage of interest, and that the analytical data accurately reflect the sample. Measurement of precision in these two requirements determines the level of confidence in the assessment. Precision is measured to identify errors and allow inferences to be made about the repeatability of an assessment. Once the precision of a method is known, the likelihood of replicating an assessment can be estimated and the level of confidence in an assessment can be characterized. More specific guidance on documenting measurement error, as well as temporal and spatial variability, is provided below.

Estimating and documenting measurement error

The process of collecting and analyzing biological data has inherent sources of variability that can obscure the discriminatory ability of an indicator. It is important to estimate effects of these sources of variability to ensure that monitoring objectives are addressed satisfactorily and so that data quality and comparability can be documented (Diamond et al. 1996, MDCB 1999). A major source of variability in biological assessments is measurement error. Measurement error is the degree to which one accurately characterizes the sampling unit or site and includes two general components: (1) natural spatial and temporal variability within the sample unit and (2) human or method errors. Natural spatial and temporal variability may lead to differences in precision or bias in an indicator that can result in inaccurate characterization of a site. Human or method errors include inconsistencies in sampling effort across sites, inappropriate use of sampling gear, inaccuracies in laboratory sorting and processing, and misidentified organisms. All of these errors can also result in mischaracterization of a site.

Human and methodological errors are controlled by using standardized and comparable methods, proper training of personnel, and quality assurance procedures (U.S. EPA 1995). Quality assurance procedures include examination of replicate field samples at some subset of the sample units (e.g., 10% of the sites) and reexamination of a proportion of samples by an independent taxonomist. For programs in which multiple field sampling crews are used, it is important to document variability in results caused by personnel. Side-by-side sampling by different field crews is done to document the magnitude of variability as a source of measurement error. Adequate training and similar experience shared across crews helps ensure that this source of error is minimized.

Documenting temporal variability among and within field seasons

It is unlikely in a monitoring program that data can be collected simultaneously from a large number of sites. Instead, sampling may be conducted over several days, weeks, or months. In many cases, indicators are implemented only during a particular season, time of day, or other window of opportunity when their signals are determined to be strong, stable, and reliable, or when stressor influences are expected to be greatest. This optimal time frame, or index period, can reduce sources of error in site characterization resulting from temporal variability (U.S. EPA 1999a). However, because an index period can span several weeks or months, it may be prudent to estimate and document variability within a field season or index period. This process is best accomplished by analyzing multiple samples, collected over time, from reference sites.

Although resource constraints often limit assessments to single index periods, it is useful to understand seasonal effects on an indicator, particularly in cases involving unexpected monitoring demands, such as spills, emergencies, and time-critical decisionmaking. Understanding the seasonal variability and expectations for biological data, using candidate reference sites, could allow data to be used for studies outside the primary index period or for other programmatic needs.

Documenting temporal variability across years

Indicator responses may change over time, even when environmental conditions remain relatively stable. Changes may be due to weather, succession, population cycles, or other natural interannual variations. Available estimates of variability across years should be examined to ensure that the indicator reflects true trends in ecological condition for characteristics that are relevant to the assessment question. To determine interannual stability of an indicator, EPA recommends that monitoring be conducted for several years at stable reference sites with minimal influence of stressors/pollutants.

Documenting spatial variability

Indicator responses to various environmental conditions must be consistent across a site class to enable reliable assessments. Locations within the reporting unit that are known to share similar ecological conditions should exhibit similar indicator results. If spatial variability occurs because of natural regional differences in physiography or habitat (e.g., elevation), it may be necessary to adjust indicator expectations and/or stratify the reporting area into more homogeneous subunits.

Use of a regional reference condition, based on an aggregate of high-quality sites, will account for "natural" spatial variability. This information is then used to determine the discriminatory capability of the indicator. Partitioning the natural variability on a spatial scale (i.e., site classification) ensures that biological response to various stressors will be similar within the site class.

5.2.7 Use of Other Types of Biological Data

Additional types of biological data may be available to, or generated by, a state for determining the status of waterbodies and for making decisions regarding impairments and the need for listings. For example, if a state shares a waterbody with another state, it must consider existing and readily available data from the state that shares the waterbody. Additional data may include fish population data (fisheries surveys, population modeling, impingement/entrainment data), endangered species data, migration data, spawning data, etc. Using these other types of biological data, states may decide to list waterbodies as impaired and initiate TMDLs to manage the impairments. When doing so, states should clearly justify the assessment and impairment decision by documenting how the biological data illustrate a violation of their water quality standards, either the designated uses, the criteria or the antidegradation policy. In many cases, additional types of biological data are interpreted as indicating violations of narrative standards or designated uses. For example, New York State listed 152 miles of the Hudson River, from the Troy Dam south, because of thermal changes from power plants leading to fish mortality occurring during power plant cooling water intake. Based on 24 years of fisheries studies, data indicated that tens to hundreds of millions of eggs, larvae, and juvenile fish of several species were killed per year by large volume, once-through cooling water users. The cumulative impact of multiple once-through cooling facilities substantially reduces the young-of-the-year population for the entire river. Data indicated that this reduction was 25%-79% for spottail shiner, 27%-63% for striped bass, 52%-60% for American shad, 44%-53% for Atlantic tomcod perch, and 33% for bay anchovy. All perennial fresh waters in New York's WQS (including the Hudson River) have a narrative standard that states these waters shall be suitable for fish propagation and survival.

Additionally, data indicating the presence of introduced, exotic, or invasive species may be used to make a use impairment decision. This is up to the state's discretion in determining whether a particular species is predominating the waterbody to the extent of impairing aesthetic or recreational uses. However, for making aquatic life use impairment decisions, the approaches outlined above will consider such species' presence in the calculation of metrics and the associate index. If the biological assessment is sufficiently robust, the impact of introduced, exotic, or invasive species will be shown by the data. The state should also be aware of any threatened and/or endangered species that may reside in or near the waterbody of concern, and may judge the use to be impaired if water quality does not support the species of concern.

5.3 How Does the State Analyze Biological Data to Determine WQS Attainment?

An important step in a bioassessment program is to analyze the data to make WQS attainment decisions and identify any impairment. The establishment of decision thresholds as benchmarks in the water quality standards of the state, or in other implementing regulations or policy or procedures documents, is key to the data analysis. This section describes the overarching strategy for analysis of biological data (5.3.1), the multimetric approach to analyzing data (5.3.2), the combination of metrics and multiple discriminant analysis (5.3.3), and a modeling approach using observed/ expected taxa (5.3.4).

State bioassessment programs should incorporate at least two key elements for analyzing bioassessment data to develop thresholds or decision criteria. These elements are index development and threshold selection. Index development can include single or multiple metrics, discriminant models, or other predictive models of the aquatic community. Thresholds are the "criteria" above which the waterbody is considered to be in attainment. The index should be developed and then verified on independent data sets. Then the attainment threshold should be established and documented. Selecting this threshold, or criterion, is perhaps the most critical element in reporting and documenting attainment status. States typically establish this threshold, and then add other thresholds to distinguish among higher (e.g., outstanding natural resource waters, excellent warmwater habitat, or excellent/good habitat) and lower assessment categories (e.g., limited resource waters, fair/poor/very poor). All thresholds, and the rationales for their selection, should be documented either in the applicable WQS, or in other implementing regulations or policies and procedures documents such as the state, territory, or authorized tribe's continuous planning process or consolidated assessment and listing methodology. More detailed descriptions of the various analytical approaches taken by states appear in Table 5-2, along with the level of information they provide. For estuaries, a different approach is used including reference thresholds for biological effects of contaminants (Long et al. 1995), sediment toxicity, and bottom dissolved oxygen (Schimmel et al. 1994).

5.3.1 Analysis of the Biological Data

Numerous methods are available for analyzing biological indicator data to assess WQS attainment status, including both univariate and multivariate analysis techniques. Bioassessment programs functioning at Levels 3 and 4 (see Table 5-1) have focused on three primary approaches. Sections 5.3.2, 5.3.3, and 5.3.4 go into more detail on each of the three approaches for states and tribes with programs of lower level of rigor to refine and enhance their existing programs.

States do not need to develop their own data analysis methods. Use of existing tools is acceptable and encouraged. Each state does need to document the specific tool it will be using (e.g., a specific multimetric index) and how it will apply this tool. Each state should document the level of information on the indicator index used (whether multimetric or discriminant/predictive model).

EPA recommends that each state establish its analytical threshold based on index values from a statistical distribution of candidate reference sites, or a discriminant model from a range of aquatic life conditions that includes reference conditions. Estimates of variance, such as a standard deviation, as well as power analysis (Fore et al. 1996) can assist in determining how many assessment levels an index may represent.

Regardless of approach, the primary purpose of an analytical threshold is to establish levels of biological quality that can be used for determining WQS attainment status in the aquatic system of interest. States need to carefully document their rationale for selecting thresholds, including thresholds that define gradations in quality or attainment status such as "good/fair/poor" or "full/partial attainment/nonattainment." The threshold should allow for relatively

Table 5-2. Description of component biological variables and predicted direction of variable response to increased perturbation

Variable	Description	Direction
Generalized core variable *Richness (assemblage)*		
1. Taxa richness	Measures the overall variety of the assemblage. Measure of biodiversity.	Decrease
2. Specific family/order richness	Number of taxa in various taxonomic families or orders that are ecologically informative. Examples are number of mayflies (Ephemeroptera) or number of darter species.	Decrease
3. Threatened and endangered species	Usually rare taxa where habitat and ecological viability at risk of depletion and ultimate extinction.	Decrease
4. Rare species	Taxa with low numbers of individuals in population.	Decrease
Composition (assemblage)		
5. Expected biota (observed/expected)	A modeled prediction of taxonomic composition in undisturbed waterbodies within natural site classes. Endpoint approaches 1 for attainment of natural or expected condition.	Decrease
6. Relative abundance	Percent composition of taxonomic groups to total number of individuals in sample, or composed within a particular taxonomic hierarchy. Examples are percent green sunfish (increases with perturbation) and percent stoneflies (Plecoptera), which decrease with perturbation.	Decrease (Increase for certain groups)
7. Compositional redundancy	Measures the change in dominance or redundancy of relative abundance as stressors increase. An example is the increase in percent dominance of one taxon as others are diminished. Evenness and diversity indexes include redundancy components	Increase
8. Keystone taxa	Targeted populations considered crucial to maintenance of assemblage or community. Example is presence of brook trout.	Decrease
9. Alien species	Taxa that are not indigenous to a particular area.	Decrease
Function (population, assemblage, or system)		
10. Reproductive success (population)	Measures some aspect of spawning and nursery success; may be representation of a variety of larval stages. Examples: young-of-year, juvenile index.	Decrease
11. Trophic structure (system)	Measures capacity of ecosystem to support primary and secondary producer/consumers. May comprise several metrics.	Decrease or Increase
12. Guilds (assemblage)	How organisms earn their living. May be trophic, habitat, feeding, or reproductive associations of organisms. May comprise several taxa.	Decrease or Increase
13. Long-lived taxa guild (assemblage)	Support of multi-year life cycle guild indicates extended good water/habitat quality. Examples are reproducing populations of trout and common abundance of stonefly nymphs.	Decrease

Table 5-2. Description of component biological variables and predicted direction of variable response to increased perturbation continued

Variable	Description	Direction
Response to stress (individual, population or assemblage)		
14. Anomalies, disease, deformities, aberrances	Sublethal effects from disease and/or toxicants: e.g., lesions, tumors, or eroded fins in fish; deformed chironomid head capsules; anomalies in striae patterns or frustule shape of diatoms.	Increase
15. Changes in regional species distribution	Range restrictions or expansions of individual species; this is typically depicted on maps comparing where a species was collected historically against current collection locales. Sensitive species typically experience range reductions; invasive aliens and tolerant species expand their ranges.	Increase
16. Other specific response signatures (individual, population or assemblage)	Any measure that has capability of diagnosing stressors. Examples include % aberrant diatoms linked to heavy metal contamination; and measures of individual health (lesions, tumors linked to toxicity).	Characteristic response
Fish-specific (variations) *Richness*		
1. Native taxa richness	Number of different native species, measure of biodiversity.	Decrease
Composition		
2. Morphological composition	Used mostly in the fish assemblage to measure affinity for water column or bottom substrate. An example is % round-bodied sucker.	Increase (?)
3. Habitat preference composition	Measures integrity of habitat to support variety of taxa. An example is number of headwater species.	Decrease
4. Genetic diversity	Genetic variation occurs even when phenotypes appear identical. The use of molecular techniques (e.g., gel electrophoresis to distinguish allozymes) are used to assess genetic diversity.	Decrease
5. Salmonid guilds	Population metrics that characterize various life stages of top carnivores.	Decrease
6. Temperature preference richness (temperature guilds)	Usually related to cold-water forms and those that are stenothermic.	Decrease
Function		
7. Specialized spawners (spawning guilds)	Excludes strategies tolerant to siltation. Measure of ability of stream reach to support a variety of reproductive strategies; affected by toxics, turbidity, sedimentation.	Decrease
8. Specialized feeders (feeding guilds)	Excluding omnivores; measure of trophic/food web complexity of fish assemblage.	Decrease
9. Biomass	Composite weight (biomass). Measure of relative productivity.	Decrease
10. Abundance	Number of individuals (abundance).	Decrease
11. Migration	Daily migrations are typically for feeding and/or predatory avoidance; most seasonal migrations are for reproduction.	Decrease

Table 5-2. Description of component biological variables and predicted direction of variable response to increased perturbation continued

Variable	Description	Direction
12. Anadromous spawning	Fish that spend most of their lives in salt water migrating to fresh water to spawn (e.g., salmon, striped bass, shad).	Decrease
13. Top carnivores	Measure of ability of food chain to support top level; affected by toxics, turbidity.	Decrease
Response to stress		
14. Morbidity	The rate of diseased or affected organisms in a specific location.	Increase
15. Tissue contamination	Measurement of pollutant(s) concentration in living organisms.	Increase
Macroinvertebrate-specific (variations) *Composition*		
1. Diversity indexes	Integrates richness and evenness in mathematical algorithm. An example is the Shannon-Wiener Diversity Index.	Decrease
2. % Dominant taxon	A specific measure of compositional redundancy found in several macroinvertebrate multimetric indices.	Increase
Function		
3. Habit representation (flow/habitat guilds)	Most preferred habit measure is % clingers, which include the insects having a fixed retreat or adaptations for attachment to surfaces in flowing water. Excludes molluscs and other non-insect taxa.	Decrease
4. Voltinism (life cycle guilds)	Measure of long-lived macroinvertebrates (univoltine, life cycles of 1 or more years) or short life cycles (multivoltine, several per year).	Increase or Decrease
Algal-specific (variations) *Composition*		
1. Diversity indexes	As described for macroinvertebrates	Decrease
2. Community similarity	Integrates richness and relative abundance for comparing the composition among sites. Adapted from Whitaker and Fairbanks. Need reference site composition or modeled composition of reference conditions, similar to O/E measure.	Decrease
Function		
3. Autecological affinity (chemical guilds)	Measures the ecological preferences of diatoms and is useful along a stressor gradient. Examples are acidobiontic, alkaliphlic, etc.	Increase or Decrease
4. Biomass	Measures indication of nutrient problem and potential for nuisance algal growth.	Increase

Source: Revised from EPA 1999a and 2000b.

straightforward decisions when biological data are compared against the thresholds to facilitate water quality management decisions. State decisions applying the threshold also need to be documented.

5.3.2 The Multimetric Approach

The most common method of data analysis is use of a multimetric index, which combines several biological variables into a single, unitless index. These variables, or metrics, are characteristics of the biota that change in some predictable way with increased human influence (Barbour et al. 1995). Use of multiple metrics to assess biological conditions maximizes the information available regarding the functions and processes of aquatic communities. For a metric to be of value, it should be (1) ecologically relevant both to the biological assemblage or community under study and to the specified program objectives, and (2) it must be sensitive to stressors (Barbour et al.1995). All metrics that fit these two criteria are potential metrics for consideration. Further analysis of this "universe" will likely eliminate some metrics because of insufficient data or because the range in data is not sufficient to distinguish between natural variability and anthropogenic effects. The analysis should identify the candidate metrics that warrant further consideration (i.e., those that are most informative).

The selected metrics can be used independently or together, depending upon the state's specific program design. A pioneer in the use of multimetric indices for bioassessment, Ohio EPA has developed indices for fish and macroinvertebrate assemblages of its streams and rivers (Yoder and Rankin 1995).

In multimetric analyses, several metrics are calculated and scored from low to high in a common scoring system. Scoring is needed because some metrics respond in different directions to anthropogenic stressors. For example, the abundance of tolerant organisms (density) increases as conditions degrade, whereas the number of intolerant taxa (richness) decreases as conditions degrade. Once the metrics have been scored using a common scale, the scores of all metrics are summed or averaged for a final index score. A multimetric index originally developed for fish assemblages in Midwestern streams (Karr 1981, Karr et al. 1986) has been adapted to streams and rivers throughout the United States and tested in lakes, reservoirs, and estuaries. Because modifications in the index may be appropriate for different regions and among waterbody types, a process for calibrating an index for ecological specificity is required. That process involves two primary steps: (1) selecting candidate metrics and testing for those that should become core metrics, and (2) developing an index by transforming metric values to unitless scores and aggregating as a multimetric index. Examples of generic metrics that are used in various water resource programs are described in Table 5-2. The response of these metrics along a biological gradient provides a means to assess condition to different levels of impairment.

Selection of metrics

Examples of ecologically relevant attributes include components of diversity, identity, composition, function, invasion by exotics, and rare and endangered species. Potential measures relevant to the ecology of the waterbody within the region or state should be evaluated.

Representative metrics from each of four primary categories should be selected: (1) *richness*, which measures for diversity or variety of the assemblage; (2) *composition*, which measures for identity and dominance; (3) *tolerance*, which measures for sensitivity to perturbation; and (4) *trophic measures*, which provide information on feeding strategies and guilds. Karr and Chu (1999) suggest that measures of individual organism health (i.e., anomalies or deformities) be used to supplement other metrics. Karr has expanded this concept to include metrics that are reflective of landscape-level attributes, thus providing a more comprehensive, multimetric approach to ecological assessment (Karr and Chu 1999).

Core metrics should be selected following initial candidate metric screening to identify those that discriminate between "good" and "poor" quality ecological conditions. Metrics that are responsive to specific pollutants or stressors, where the response is well characterized, are most useful as diagnostic tools. Core metrics should be selected to represent diverse aspects of structure, composition, individual health, or processes of the aquatic biota. Together they form the foundation for a sound, integrated analysis of the biotic condition to judge attainment of biological criteria or designated aquatic life uses. The ability of a metric to discriminate between reference conditions and stressed conditions (determined by abiotic, or nonbiological, judgement criteria) is crucial to selecting core metrics. Multiple metrics should be selected to provide a strong and predictable relationship with biological conditions.

Combining metrics into an index

Two basic approaches are used to develop metric expectations and scoring criteria as a basis for index development (Simon and Lyons 1995). The approaches are to use data from reference sites (i.e., composited reference condition) or data from sites representing a range of conditions (i.e., a disturbance gradient). If reference sites are used, there should be a sufficient number of reference sites and samples available to define reference conditions. If data from sites representing a range of conditions (disturbance gradient) are used, they should reflect the entire range of abiotic influence, from minimal human influence to degradation. In either case, a regional reference condition should be developed for each site class (typically termed a bioregion).

Metrics vary in their scale; they may be integers, percentages, or dimensionless numbers. Prior to developing an integrated index for assessing biological condition, it a state should standardize core metrics via a transformation to unitless scores. Recent research has shown that transforming metric values into unitless scores is best done on a numerical scale from 100 to 0 (Hughes et al. 1998, U.S. EPA 1999a). Under such an approach, the data from all sites for each metric, including reference sites, are truncated to the 95th percentile to prevent outliers and extreme values from adversely influencing scoring criteria. (Note: For those metrics that tend to increase in value as the disturbance gradient increases, the 5th percentile is used.) The range from the 95th percentile to the minimum possible value is then subdivided from 100 to 0, with 100 being the maximum score. Finally, the summation of all metric scores is averaged to provide a 100-point scale for the index.

An index provides a means of integrating information from a composite of biological metrics. Aggregation of metrics simplifies management and decisionmaking so that a single-index value is used to determine whether action is needed. The common elements in the development of any analytical assessment tool are use of (1) an initial data set to develop (calibrate) the index and (2) a confirmation data set to test (validate) the index. The initial and confirmation data may be from the same set of biological data, randomly divided, or they may be from two consecutive years of biological data used separately. All sites in each data set are identified by degradation class (e.g., reference versus stressed). To avoid circularity, identification of reference and stressed classes should be made based on nonbiological (abiotic) information, such as the quality of the riparian zone and other habitat features, the presence of known discharges and nonpoint sources, the extent of impervious surface in the watershed, and the extent of land use practices, among other indicators.

Analytical threshold

The population of reference sites normally is used to determine the threshold that separates acceptable from unacceptable biological condition. Reference can also be used to refine aquatic life designated uses by clearly defining the level of biological condition associated with each use as discussed earlier in Section 5.1.1. A population statistic, such as the 25th percentile (Yoder and Rankin 1995, DeShon 1995, Barbour et al. 1996) or 10th percentile (Roth et al. 1997) of the reference sites is a commonly used threshold for multimetric indices. A 25th or 10th percentile is used to recognize that conditions at candidate reference sites are variable, and those at the lower end of the reference scale have a certain level of uncertainty in their quality. This recognition does not mean that 25% of the candidate reference sites are impaired, but that these sites may need closer scrutiny or investigation to assess their condition. The greater the uncertainty in accurately selecting true reference sites, the higher the threshold percentile should be. In addition, precision estimates of the bioassessment methods provide a range of values in which a site condition may not be assessed confidently as either acceptable or unacceptable. In such case, more investigation may be warranted.

5.3.3 Combining Metrics and Multiple Discriminant Analysis

A variety of approaches can be used to combine metrics for an attainment determination. Maine DEP employs a hierarchical decisionmaking technique, which is an example of a discriminant model that uses a variety of biological metrics. It begins with statistical models (linear discriminant analysis) to make an initial prediction of the classification of an unknown sample by comparing it with characteristics of each class identified in the baseline database (Davies et al. 1993). The output of the primary statistical model is a list of probabilities of membership for each of four groups designated as classes A (the highest aquatic life use), B, C, and nonattainment (NA) of Class C. All sites are given an *a priori* aquatic life use of A, B, or C based on waterbody uses and administrative decisions. Stream biologists from Maine DEP assigned a training set of streams to form aquatic life use classes and tested the argument with water chemistry data (see Davies et al. 1993 for description of how ALUS classes were established). Subsequent models are two-way discriminant models to distinguish between a given class and any higher classes as one group, and any lower classes as a second group. The

model uses 31 quantitative measures of community structure, including the Hilsenhoff Biotic Index, Generic Species Richness, EPT, and EP values. Monitored test sites are then assigned to one of the four classes based on the probability of that result, and uncertainty is expressed for intermediate sites. The classification can be the basis for management action if a site does not meet its designated use (A, B, or C) or the basis for reclassification to a higher class if the site has improved.

Analytical threshold

The Maine DEP discriminant models predict the membership of a site in one of Maine's aquatic life use classes A, B, or C, or nonattainment (NA). Assignment to a single class must be based on a probability predicted by the submodel of 0.6 or greater. If the model indicates a site is actually in a lower biological class than its designated legislative class, then the site is not attaining its aquatic life use (e.g., the site is listed as Class B, but the discriminant analysis assigns the biota to Class C). If the model fails to assign a class by the required probability, best professional judgment is used.

5.3.4 Modeling Approach Using Observed/Expected Taxa

Another approach, which is used in Oregon and extensively by the U.S. Forest Service, is based on an empirical (statistical) discriminant function model that predicts the aquatic macroinvertebrate fauna that would be expected to occur at a site in the absence of environmental stress (Simpson et al. 1996). A comparison of the invertebrates predicted to occur at the test sites with those actually collected provides a measure of biological impairment at the tested sites. The predicted taxa list also provides a "target" description of the invertebrate community to measure the success of restoration measures. The type of taxa predicted by the model also may provide clues as to the type of impact a sampled site is experiencing. This information can be used to facilitate further investigations and design control/restoration measures. The models are based on a stepwise progression of multivariate and univariate analyses and have been developed for several regions and various habitat types found in lotic systems. Each model is tailored to specific regions (or states) to provide the most accurate predictions for the seasonal and habitat sampled. (See Hawkins et al. 2000b for a more complete description of how this is done.) This approach is being evaluated by EPA. States using this observed/expected approach will need to describe in their methodologies how their model was built and tested for waterbodies.

Analytical threshold

Oregon combines metrics and multivariate models to assess biological condition. In deciding to list or delist impaired waters, Oregon considers aquatic communities (primarily macroinvertebrates) to be impaired if they are found to be at 60% or less of the expected reference community for both multimetric scores and multivariate model scores. Streams with either multimetric scores or multivariate scores between 61% and 75% of expected reference communities are considered to be "streams of concern." Streams with greater than 75% of

expected reference communities using either multimetric or multivariate models are considered unimpaired.

5.3.5 Determining Water Quality Standards Attainment

As stated in section 5.1, biological assessment data are important for measuring the attainment of water quality standards for the protection of aquatic life. Biological assessments reflect the total cumulative impact of all stressors over a period of time on a waterbody using the biological community as an indicator. In order for States to best use biological assessment data when determining water quality standards attainment, States should either define their Aquatic Life Uses in their WQS in terms of the expected biological condition for that class and type of waterbody, adopt numeric biocriteria in their WQS, or evaluate the bioassessment data pursuant to well-described implementation procedures or translator mechanisms that define quantitative thresholds that are described either in the WQS or alternatively in other implementing regulations or policies and procedures documents such as the state, territory, or authorized tribe's continuous planning process or consolidated assessment and listing methodology. Each of these approaches should have adequate documentation in the assessment and listing methodology of how the data will be used when addressing all the key elements of a State biological assessment program. Additionally, this documentation should include caveats relating to the known quality and rigor of the data which has been documented earlier in Table 5.1 and Section 5.2.

Although biological data and biological standards can be used to identify water quality impairments, biological data alone, does not usually identify the causes of impairments. Identification of the causes of biological impairments usually requires evaluation of the biological data and other information on watershed conditions. The state, territory, or authorized tribe's assessmesnta nd listing methodology should describe how biological data will be used to determine the cause of an impairment and whether a use is impaired by a pollutant, if this has not already been established in the WQS or other implementing policy or procedure document. For guidance on procedures for identifying causes of biological impairment, see the Stressor Identification guidance document (U.S. EPA 2001).

5.4 References

American Public Health Association (APHA). 1992. Standard methods for the examination of water and wastewater. 18th ed. American Public Health Association, American Water Works Association, and Water Pollution Control Federation. Washington, DC.

Bahls L. 1993. Periphyton bioassessment methods for Montana streams. Water Quality Bureau, Department of Health and Environmental Science, Helena, MT.

Bailey RC, Norris RH, Reynoldson TB. 2001. Taxonomic resolution of benthic macroinvertebrate communities in bioassessments. J N Am Benthol Soc 20(2):280-286.

Barbour MT, Gerritsen J. 1996. Subsampling of benthic samples: A defense of the fixed organism method. J N Am Benthol Soc 15:386-392.

Barbour MT, Gerritsen J, Griffith GE, Frydenborg R, McCarron E, White JS, Bastian ML. 1996. A framework for biological criteria for Florida streams using benthic macroinvertebrates. J N Am Benthol Soc 15:185-211.

Barbour MT, Stribling JB, Karr JR. 1995. The multimetric approach for establishing biocriteria and measuring biological condition. In: Davis W, Simon T, eds. Biological assessment and criteria: Tools for water resource planning and decisionmaking. Ann Arbor, MI: Lewis Publishers, pp. 63-76.

Bauer SB, Burton TA. 1993. Monitoring protocols to evaluate water quality effects of grazing management on western rangeland streams. U.S. Environmental Protection Agency, Region 10. Seattle, WA. EPA 910-R-93-017.

Caton LW. 1991. Improved subsampling methods for the EPA "rapid bioassessment" benthic protocols. Bull N Am Benthol Soc 8(3):317–319.

Courtemanch DL. 1996. Commentary on the subsampling procedures used for rapid bioassessments. J N Am Benthol Soc 15:381–385.

Cuffney TF, Moulton SR, Carter JL, Short TM. 2000. Abstract. Fixed-count and proportional benthic invertebrate subsampling methods: A comparison of efficacy and cost. Bull N Am Benthol Soc 17(1):144.

Davies SP, Tsomides L, Courtemanch DL, Drummond F. 1993. Maine Biological Monitoring and Biocriteria Development Program. Maine Department of Environmental Protection, Bureau of Water Quality Control, Division of Environmental Evaluation and Lake Studies. Augusta, ME.

DeShon JE. 1995. Development and application of the invertebrate community index (ICI). In: Davis WS, Simon TP, eds. Biological assessment and criteria: Tools for water resource planning and decision making. Boca Raton, FL: Lewis Publishers, pp. 217–243.

Diamond JM, Barbour MT, Stribling JB. 1996. Characterizing and comparing bioassessment methods and their results: A perspective. J N Am Benthol Soc 15:713-727.

Dixit SS, Smol JP, Kingston JC, Charles DF. 1992. Diatoms: Powerful indicators of environmental change. Environ Sci Technol 26:23–33.

Engle VD. 2000. Application of the indicator evaluation guidelines to an index of benthic condition for Gulf of Mexico estuaries. Chapter 3. In: Jackson L, Kutz J, Fisher W, eds. Evaluation guidelines for ecological indicators. U.S. Environmental Protection Agency, Office of Research and Development, Research Triangle Park, NC. EPA 620-R-99-005.

Ferguson RL, Wood LL. 1994. Rooted vascular aquatic beds in the Albemarle-Pamlico estuarine system. National Marine Fisheries Service, Beaufort, NC. Project No. 94-02.

Florida Department of Environmental Protection (FL DEP). 1996. Standard operating procedures for biological assessment. Florida Department of Environmental Protection, Biology Section. July 1996.

Fore LS, Karr JR, Wisseman RW. 1996. Assessing invertebrate responses to human activities: Evaluating alternative approaches. J N Am Benthol Soc 15(2):212-231.

Gerritsen J, Barbour MT, King K. 2000. Apples, oranges, and ecoregions: On determining pattern in aquatic assemblages. J N Am Benthol Soc 19:487-496.

Gerritsen J, White J, Barbour MT. 1996. Variability of Wyoming stream habitat assessment and biological sampling. Prepared for Wyoming Department of Environmental Quality, Sheridan, WY.

Hawkins CP, Norris RH, Gerritsen J, Hughes RM, Jackson SK, Johnson RK, Stevenson RJ. 2000a. Evaluation of the use of landscape classifications for the prediction of freshwater biota: Synthesis and recommendations. J N Am Benthol Soc 19(3):541-556.

Hawkins CP, Norris RH, Hogue JN, Feminella JW. 2000b. Development and evaluation of predictive models for measuring the biological integrity of streams. Ecol Appl 10:1456-1477.

Hawkins CP, Norris RH. 2000. Effects of taxonomic resolution and use of subsets of the fauna on the performance of RIVPACS-type models. In: Wright JF, Sutcliffe DW, Furse MT, eds. Assessing the biological quality of fresh waters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, UK. pp. 217-228.

Hayslip GA. 1993. EPA Region 10 in-stream biological monitoring handbook (for wadable streams in the Pacific Northwest). Region 10, U.S. Environmental Protection Agency, Environmental Services Division, Seattle, WA. EPA 910-9-92-013.

Hilsenhoff WL. 1987. An improved biotic index of organic stream pollution. Great Lakes Entomol 20:31-39.

Holland AF, ed. 1990. Near coastal program plan for 1990: Estuaries. Office of Research and Development, U.S. Environmental Protection Agency, Narragansett, RI. EPA 600-4-90-033.

Hughes RM, Kaufmann PR, Herlihy AT, Kincaid TM, Reynolds L, Larsen DP. 1998. A process for developing and evaluating indices of fish assemblage integrity. Can J Fish Aquat Sci 55:1618-1631.

Idaho Department of Environmental Quality (IDEQ). 1999. Draft V.2, Idaho rivers ecological assessment framework. IDEQ River Bioassessment Team.

Intergovernmental Task Force on Monitoring Water Quality (ITFM). 1995. The strategy for improving water-quality monitoring in the United States: Final report of the Intergovernmental Task Force on Monitoring Water Quality. U.S. Geological Survey, Reston, VA.

Karr JR. 1981. Assessment of biotic integrity using fish communities. Fisheries 66:21-27.

Karr JR, Chu EW. 1999. Restoring life in running waters: Better biological monitoring. Washington, DC: Island Press.

Karr JR, Fausch KD, Angermeier PL, Yant PR, Schlosser IJ. 1986. Assessment of biological integrity in running waters: A method and its rationale. Special Publication 5. Illinois Natural History Survey, Champaign, IL.

Kentucky Department of Environmental Protection. 1993. Methods for assessing biological integrity of surface waters. Division of Water, Kentucky Department of Environmental Protection, Frankfort, KY.

Kerans BL, Karr JR, Ahlstedt SA. 1992. Aquatic invertebrate assemblages: Spatial and temporal differences among sampling protocols. J N Am Benthol Soc 11:377-390.

Long ER, MacDonald DD, Smith SL, Calder FD. 1995. Incidence of adverse biological effects within ranges of chemical concentrations in marine and estuarine sediments. Environ Manage 19:81-97.

Lonzarich D. 1994. Stream fish communities in Washington: Patterns and processes. PhD dissertation, University of Washington.

McCormick FH, Peck DV. 2000. Application of the indicator evaluation guidelines to a multimetric indicator of ecological condition based on stream fish assemblage. Chapter 4. In: Jackson L, Kurtz J, Fisher W, eds. Evaluation guidelines for ecological indicators. U.S. Environmental Protection Agency, Office of Research and Development, Research Triangle Park, NC. p. 107. EPA 620-R-99-005.

Methods and Data Comparability Board (MDCB). 1999. Towards a definition of a performance-based approach to laboratory methods. Version 5.2. In: http://wwwdwimdn.er.usgs.gov/ pmethods.

Oklahoma Conservation Commission (OCC). 1993. Development of rapid bioassessment protocols for Oklahoma utilizing characteristics of the diatom community. Oklahoma Conservation Commission, Oklahoma City, OK.

Orth RJ, Nowak JF, Anderson GF, Whiting JR. 1993. Distribution of submerged aquatic vegetation in the Chesapeake Bay and its tributaries and Chincoteague Bay—1992. Prepared by Virginia Institute of Marine Science, Gloucester Point, Virginia for the U.S. Environmental Protection Agency, Chesapeake Bay Program Office, Annapolis, MD.

Resh VH, McElvary EP. 1993. Contemporary quantitative approaches to biomonitoring using benthic macroinvertebrates. In: Rosenberg DM, Resh VH, eds. Freshwater biomonitoring and benthic macroinvertebrates. New York: Chapman and Hall, pp. 159-194.

Rosen BH. 1995. Use of periphyton in the development of biocriteria. In: Davis WS, Simon TP, eds. Biological assessment and criteria. Tools for water resource planning and decision making. Boca Raton, FL: Lewis Publishers, pp. 209-215.

Roth NE, Southerland MT, Chaillou JC, Volstad JH, Weisberg SB, Wilson HT, Heimbuch DG, Seibel JC. 1997. Maryland biological stream survey: Ecological status of non-tidal streams in six basins sampled in 1995. Report no. CBWP-MANTA-EA-97-2. Maryland Department of Natural Resources, Annapolis, MD.

Rott E. 1991. Methodological aspects and perspectives in the use of periphyton for monitoring and protecting rivers. In: Whitton BA, Rott E, Friedrich G, eds. Use of algae for monitoring rivers. Institut fur Botanik, University of Innsbruck, Austria.

Schimmel SC, Melzian BD, Campbell DE, Stubel CJ, Benyi SJ, Rosen JS, Buffum HW. 1994. Statistical summary: EMAP-Estuaries Virginian Province—1991. Office of Research and Development, U.S. Environmental Protection Agency, Narragansett, RI. EPA 620-R-94-005.

Simon TP, Lyons J. 1995. Application of the index of biotic integrity to evaluate water resource integrity in freshwater ecosystems. In: Davis WS, Simon TP, eds. Biological assessment and criteria. Tools for water resource planning and decision making. Boca Raton, FL: Lewis Publishers, pp. 245-262.

Simpson J, Norris R, Barmuta L, Blackman P. 1996. Australian River assessment system: National river health program predictive model manual. http.//ausrivas.canberra.au.

Spindler P. 1996. Using ecoregions for explaining macroinvertebrate community distribution among reference sites in Arizona. 1992. Arizona Department of Environmental Quality, Hydrologic Support and Assessment Section, Flagstaff, AZ.

U.S. Environmental Protection Agency (U.S. EPA). 1989. Rapid bioassessment protocols for use in streams and rivers. Benthic macroinvertebrates and fish. Plafkin JL, Barbour MT, Porter KD, Gross SK, Hughes RM. Office of Water Regulations and Standards, Washington, DC. EPA 440-4-89-001.

U.S. EPA. 1990a. Biological criteria: National program guidance for surface waters. Office of Water Regulations and Standards, Washington, DC. EPA 440-5-90-004.

U.S. EPA. 1991a. Policy on the use of biological assessments and criteria in the water quality program. Attachment A of Memorandum from Tudor Davies, Director. Office of Science and Technology to Water Management Division Directors, Regions I–X.

U.S. EPA. 1992. Framework for ecological risk assessment. Washington, DC. EPA 630-R-92-001.

U.S. EPA. 1993a. EPA region 10 in-stream biological monitoring handbook (for wadable streams in the Pacific Northwest). Gretchen Hayslip. Region 10, Environmental Services Division, Seattle, WA. EPA 910-9-92-013.

U.S. EPA. 1995. Generic quality assurance project plan guidance for programs using community-level biological assessment in wadable streams and rivers. Office of Water, Washington, DC. EPA 841-B-95-004.

U.S. EPA. 1996a. Biological criteria: Technical guidance for streams and small rivers. Gibson G, Barbour M, Stribling J, Gerritsen J, Karr J. Office of Science and Technology, Health and Ecological Criteria Division, Washington, DC. EPA 822-B-96-001.

U.S. EPA. 1996b. Summary of state biological assessment programs for streams and rivers. Davis WS, Snyder BD, Stribling JB, Stoughton C. Office of Planning, Policy, and Evaluation, Washington, DC. EPA 230-R-96-007.

U.S. EPA. 1998. Lake and reservoir bioassessment and biocriteria Gerritsen J, Carlson R, Charles DL, Dycus D, Faulkner C, Gibson GR, Kennedy RH, Markowitz SA. Technical guidance document. Office of Water, Washington, DC. EPA 841-B-98-007.

U.S. EPA. 1999a. Rapid bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates and fish. 2nd ed. Barbour MT, Gerritsen J, Snyder BD, Stribling JB. Office of Water, Washington, DC. EPA 841-B-99-002.

U.S. EPA. 1999b. Quantifying physical habitat in wadeable streams. Kaufmann PR, Levine P, Robison EG, Seeliger C, Peck DV. Office of Research and Development, Washington, DC. EPA/620/R-99/003.

U.S. EPA. 2000a. Estuarine and coastal marine waters: Bioassessment and biocriteria technical guidance. Gibson GR, Bowman ML, Gerritsen J, Snyder BD. Office of Water, Washington, DC. EPA 822-B-00-024.

U.S. EPA. 2000b. Evaluation guidelines for ecological indicators. Jackson LE, Kurtz JC, Fisher WS. Office of Research and Development, Research Triangle Park, NC. EPA 620-R-99-005.

U.S. EPA. 2001. Stressor identification guidance document. Office of Water, Washington, DC. EPA 822-B-00-025.

U.S. EPA. 2002a. Methods for evaluating wetland condition: Wetlands classification. Office of Water; Washington, DC. EPA 822-R-02-017.

U.S. EPA. 2002b. Methods for evaluating wetland condition: Using vegetation to assess environmental conditions in wetlands. Office of Water; Washington, DC. EPA 822-R-02-020.

Yoder CO, Rankin ET. 1995. Biological criteria program development and implementation in Ohio. In: Davis WS, Simon TP, eds. Biological assessment and criteria: Tools for water resource planning and decision making. Boca Raton, FL: Lewis Publishers, pp. 109-144.

6. Using Toxicity Data as Indicators of Water Quality

Contents

6. Using Toxicity Data as Indicators of Water Quality

The whole-effluent approach to toxics control for the protection of aquatic life involves the evaluation of substances using acute and chronic tests to measure the toxicity of wastewater and ambient waters. Whole-effluent toxicity (WET) testing is an important component of the U.S. Environmental Protection Agency's (EPA's) integrated approach for controlling the discharge of toxic chemicals and other materials into surface waters. Such WET tests are typically conducted in concert with other types of monitoring such as chemical, physical, and biological assessments. Toxicity is a valuable indicator for assessing and protecting against impacts on water quality and designated uses caused by the aggregate toxic effect of pollutants. Like chemical-specific limitations and standards, WET tests provide some predictive capability to assess the occurrence of toxicity under predefined conditions. Instream evaluation of populations of aquatic species (bioassessments) provides information on past exposures of organisms to toxic conditions, but this approach only estimates reactions to exposures that have already occurred. Instream biomonitoring procedures are therefore always reactive to an insult rather than predictive and protective. Contaminants may flow directly from industrial and municipal waste dischargers, may come from polluted runoff in urban and agricultural areas, or may collect in the sediments. Toxicity evaluations can be used to assess the type and extent of degraded water quality.

Toxicity tests used for whole-effluent and surface waters include surrogate freshwater or marine (depending on the mixture of effluent and receiving water) plants, invertebrates, and vertebrates (U.S. EPA 1991, EPA/505/2-001). EPA has published CFR part 136 methods for WET tests (60 Fed Reg 53529, Oct. 16, 1995) and the manuals are incorporated by reference, so in effect the methods are regulations. These methods are contained in the following three documents:

- Methods for Measuring the Acute Toxicity of Effluents and Receiving Waters to Freshwater and Marine Organisms (EPA/600/4-90/027F) (U.S. EPA 1993)

- Short-Term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Water to Freshwater Organisms (EPA/600/4-91/002) (U.S. EPA 1994a)

- Short-Term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Water to Marine and Estuarine Organisms (EPA/600/4-91/003) (U.S. EPA 1994b)

In WET tests, aquatic organisms (plants, invertebrates, and vertebrates) are exposed to samples of effluent in the laboratory. These exposures are conducted under controlled conditions and the response(s) of the organisms recorded. Acute toxicity is generally measured using a multiconcentration test, termed a definitive test, that consists of a control (clean water) and several (generally five) effluent concentrations. The tests are designed to provide concentration-response information, expressed as the percentage of effluent concentration that is lethal to 50% of the test organisms (LC50) within the prescribed period. The tests can also be used to determine the highest effluent concentration at which survival is not statistically significantly different from the control. Acute tests (U.S. EPA 1993, 1999) generally use death as the measured effect of a given effluent over 24 to 96 hours. Sublethal WET tests (often described as chronic tests) use longer durations of exposure (up to 9 days) to ascertain the adverse effects of an effluent on survival, growth, and/or reproduction of the organisms. For freshwater ecosystems, EPA has focused on short-term tests for three species, designed to estimate the

chronic toxicity in a water sample (U.S. EPA 1994a, 1999). These methods include a fish, larval fathead minnow (*Pimephales promelas*), a zooplankton (*Ceriodaphnia dubia*), and an alga (*Selenastrum capricornutum*). The marine and estuarine short-term tests estimate chronic toxicity (U.S. EPA 1994b, 1999) with two fish species, sheepshead minnow (*Cyprinodon variegatus*) and the inland silverside (*Menidia berylina)*, a red alga (*Champia parvula*), an East Coast mysid (*Mysidopsis bahia*), and a sea urchin (*Arbacia punctulata*). The EPA toxicity tests and other single-species tests were intended to be screening tools (i.e., to indicate the potential for wastewater or ambient water samples to cause biological community impacts, characterizing relative ecosystem effects) and "early warning" signals (a measurement that indicates the potential for aquatic ecosystem impairment prior to actual damage to biological communities (U.S. EPA 1991, 1994a). The toxicity tests are applicable to ambient water samples regardless of the sources (i.e., point or nonpoint) of contaminants. The results of all tests may be used to make quantitative estimates of the degree of toxicity of the test material.

Sediment contamination is a widespread environmental problem that can pose a threat to aquatic ecosystems. Sediment acts as a reservoir for common chemicals such as pesticides, herbicides, polychlorinated biphenyls (PCBs), polycyclic aromatic hydrocarbons (PAHs), and metals such as lead, mercury, and arsenic. Contaminated sediments may be directly toxic to aquatic life (organisms found in the water and in or near the sediment) or can be a source of contaminants for bioaccumulation (where a substance is taken up by an organism) in the food chain. Protecting sediment quality is an important part of restoring and maintaining the biological integrity of our nation's waters. Sediment is an integral component of aquatic ecosystems, providing habitat, feeding, spawning, and rearing areas for many aquatic organisms. Because sediment serves as a reservoir for contaminants, it is a source of contaminants to the water column and organisms. The extent and severity of sediment contamination in the United States, as documented in the National Sediment Inventory (U.S. EPA, 1997) and through contaminated site histories, emphasize the need for better tools for reducing and preventing sediment contamination.

Whole-sediment toxicity tests are an important tool for sediment quality assessment. They directly measure sediment toxicity to a test species under laboratory conditions, and are especially valuable because they account for interactive effects of chemical mixtures. Benthic community analyses are also useful for sediment assessment, because they account for instream conditions. Sediment toxicity testing should be conducted to characterize the nature and extent of contamination.

EPA has published the following guidance documents, which provide laboratory methods for measuring the toxicity of whole sediments:

- Methods for Assessing the Toxicity of Sediment-Associated Contaminants with Estuarine and Marine Amphipods (U.S. EPA 1994c)

- Methods for Measuring the Toxicity and Bioaccumulation of Sediment-Associated Contaminants with Freshwater Invertebrates, second edition (U.S. EPA 2000)

- Methods for Assessing the Chronic Toxicity of Marine and Estuarine Sediment-Associated Contaminants with the Amphipod *Leptocheirus plumulosus*, March 2001. EPA-823-F-01-008

When the effluent toxicity tests, receiving water toxicity tests, and whole-sediment toxicity tests are effectively used as the primary water quality indicator, the sample locations can be further characterized by the use of Toxicity Identification Evaluations (TIEs) to help identify the causative pollutants or pollutant categories.

First, appropriate effluent, receiving water, and sediment toxicity tests should be performed. EPA provides guidance on the sample size to allow for concurrent or subsequent chemical testing. Testing should be conducted to characterize the nature and extent of contamination, with the appropriate lethal and sublethal toxicity tests. There may be limitations in relying on whole-effluent and/or sediment toxicity tests to determine attainment of water quality standards (WQS). Because test organisms are selected on the basis of overall sensitivity to chemical pollutants and ecological relevance, toxicity tests are conducted with a limited range of species whose sensitivity may not be known when the chemicals of concern are unknown. On the other hand, toxicity tests have strengths in being able to detect effects from unknown or unmeasured chemicals and interactive toxicity of multiple chemicals.

6.1 How Are Toxicity Data Used Within the Context of the State's WQS?

Typically, toxicity data are used to interpret a state's narrative WQS of "no toxics in toxic amounts." Narrative criteria can be the basis for limiting toxicity in waste discharges where a specific pollutant can be identified as causing or contributing to the toxicity but there are no numeric criteria in the state standards, or where toxicity cannot be traced to a particular pollutant. Section 131.11(a)(2) requires states to develop implementation procedures that explain how the state will ensure that narrative criteria are met.

The state should describe what types of toxicity data it is requiring or considering requiring, and the appropriate test methods. WET data are generated when effluent samples are tested. Tests with receiving water samples and ambient water tests using the promulgated WET test methods should also be documented. Typically, tests with effluent are conducted with a specific discharger's effluent, using either ambient or standard laboratory water as dilution water. Receiving water toxicity tests are conducted with ambient water alone, compared with a standard laboratory dilution water, and may be done with a dilution series effluent test. Sediment toxicity test data may be required to assess contaminated sites.

Some states, however, have adopted numeric criteria for WET. EPA regulations (40 CFR 122.44) cover the national surface water toxics control program. These regulations are linked to WQS requirements and specifically address the control of pollutants with and without numeric criteria. For example, section 122.44 (d)(1)(v) provides the permitting authority with several options for establishing effluent limits when a state does not have chemical-specific numeric criteria for a pollutant at a concentration that causes or contributes to a violation of the narrative criteria. Where a state, territory, or authorized tribe adopts narrative criteria for toxic pollutants

to protect designated uses, the state, territory or authorized tribe must provide information identifying the method by which it intends to regulate point source discharges of toxic pollutants on water quality limited segments based on such narrative criteria. Such information may be included as part of the standards or may be included in documents generated by the state, territory, or authorized tribe in response to the Water Quality Planning and Management Regulations (40 CFR part 35). If a state standard is not in attainment, the water should be listed. Information regarding how the state, territory or authorized tribe intneds to regulate point source discharges of toxic pollutants on water quality limited segments based on narrative criteria is discussed in a December 1988 EPA policy and guidance, Guidance for State Implementation of Water Quality Standards for CWA section 303(c)(2)(B).

The regulatory basis for requiring WET is found in EPA regulations at 40 CFR 122.44(d)(1)(v). These regulations require NPDES permits to contain WET limits where a permittee has been shown to cause, have the reasonable potential to cause, or contribute to an instream excursion of a narrative criterion. State implementation procedures should, at a minimum, specify or reference methods to be used in implementing WET controls.

The choice of species depends on the type of regulatory WET test to be conducted and the most representative test species (e.g., for a freshwater vs. marine discharger); the test method manuals stipulate species-specific test conditions. Effluent samples are typically collected as 24-hour composite samples and tested in static renewal or static toxicity tests. The type and duration of the tests and species required must be documented. Procedures for the collection of effluent and ambient water samples, selection of the test method, and species to use under different testing conditions are described in the three promulgated WET methods (U.S. EPA 1993, 1994a,b). Similarly, the sediment tests are described in guidance (U.S. EPA 1994c, 2000, 2001).

6.2 What Actions Does the State Take To Assess and Document Data Quality?

If the state is using WET data, receiving water data, and sediment toxicity test data, it should document the relevant QA/QC procedures to document data quality. The manuals that provide the methods for the promulgated WET test methods (U.S. EPA 1993, Methods for Measuring the Acute Toxicity of Effluents and Receiving Water to Freshwater and Marine Organisms; U.S. EPA 1994a, Short-Term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Water to Marine and Estuarine Organisms; and U.S. EPA 1994b, Short-Term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Water to Freshwater Organisms) describe the QA/QC specific to these WET test methods. Specifically, Chapter 4 of each of the test manuals contains the QA/QC specifications for the promulgated WET test methods. When the WET test methods are used, Chapter 4 (Quality Assurance) should be strictly adhered to in order to ensure that the outcome of the test results is of the best quality.

EPA has developed sediment methods for aquatic invertebrates (U.S. EPA 2000, Methods for Measuring the Toxicity and Bioaccumulation of Sediment-Associated Contaminants with Freshwater Invertebrates; U.S. EPA 1994c, Methods for Assessing the Toxicity of Sediment-Associated Contaminants with Estuarine and Marine Amphipods; and U.S. EPA 2001, Methods for Assessing the Chronic Toxicity of Marine and Estuarine Sediment-Associated Contaminants

with the Amphipod *Leptocheirus plumulosus* - First Edition, EPA/600/R-01/020). These manuals also have QA guidance for obtaining organisms, water, and sediment samples, and conducting the toxicity test.

6.2.1 How Does the State Define Data Quality?

EPA encourages states, territories, interstate commissions, and authorized tribes to use the data quality objectives (DQO) process to define minimum-quality data requirements. The DQO process is a systematic procedure for defining the criteria that the data collection design should satisfy, including when to collect samples, where to collect samples, how many samples to collect, and the tolerable level of decision error for the project.

States should also document required procedures that ensure the quality of the data collected. These include information on sample collection and handling protocols, use of standardized methods, QC procedures, and data management. These procedures are generally included in a QA project plan or SOP. It is becoming increasingly clear that these procedures should be available to other organizations such as tribal, interstate, state, Federal, academic, and volunteer citizen groups that also monitor water quality. These stakeholders may agree to meet state data quality requirements if the agency clearly spells out these requirements in its assessment and listing methodology or other readily available and well-publicized documents.

6.2.2 How Does the State Review and Evaluate Data Quality?

The term “data quality assessment” means the scientific and statistical evaluation to determine if data obtained from toxicity monitoring operations are appropriate and of sufficient quality and quantity to support water quality attainment decisions. As discussed in Chapter 3, the user must know in what context a data set is to be used in order to determine whether the data set is adequate. Guidance for assessing the quality of available data sets is provided in detail in Practical Methods for Data Quality Assessment (EPA/600/R-96/084).

For assessing whether toxicity data are acceptable, EPA recommends a tiered approach (as described in Chapter 3). In this approach, the first step is to screen all reports to determine whether appropriate procedures were used and QA/QC measures were in place (e.g., whether SOPs were in place describing each step). The second step is to evaluate whether samples were collected under the appropriate conditions. The next step is to review the sample collection and any associated analytical methods to determine compatibility with the agency’s QA/QC requirements and SOPs. Then the evaluator should determine whether the sample collection and analytical methods were actually followed in creation of the data set. The final step is to assess whether the metadata accompanying the data set meet the agency’s requirements.

Upon determining that the data meet basic documentation requirements, the evaluator should decide whether additional screening of the actual data sets is needed. A state may consider reviewing reference toxicant values outside the control chart values and dose-response data for each test. If the data reveal potential problems or errors in the collection or analysis, an in-depth analysis of QA/QC procedures may be appropriate. This screening could include reviews of

QA/QC reports to determine whether the data set meets QA/QC requirements regarding measurement systems, the approach to handling missing data and problem data sets, or any deviations from SOPs.

6.2.3 How Does the State Document the Quality of the Data Used To Support WQS Attainment Decisions?

The 305(b) Consistency Workgroup developed a table to assist in documenting the quality of toxicity data and information used to support WQS attainment decisions (U.S. EPA 1997). This table is reproduced as Table 6-1.

6.3 How Does the State Analyze and Interpret Toxicity Data To Determine WQS Attainment/Impairment?

A state should document the methods used to analyze effluent, receiving water, and sediment toxicity test methods. The state's procedures should be specific about the method used; that is, LC50s must be calculated following the methods in the promulgated WET manuals. The state should also define the control used and the receiving water or reference site sediment and how significantly different comparisons were determined.

The most important element of the state, territory or authorized tribe's assessment and listing methodology is documentation of how the state analyzes and interprets data to determine WQS attainment and identify impaired waters. This documentation should be consistent with the state's, territory's, or authorized tribe's implementation procedures that are described either in the WQS or alternatively in other implementing regulations or policies and procedures documents such as the continuous planning process or consolidated assessment and listing methodology. An assessment methodology should take into account the balance between desired minimum data requirements from a strict scientific perspective and the practical realities of the availability of information and the strength of the available evidence. For example, a state's methodology could require a minimum level of decision errors for making an attainment decision, except in cases where overwhelming evidence of impairment is found. An example of overwhelming evidence would be repeated sampling events showing very high toxicity values for each species, regardless of the known cause of the toxicity.

6.3.1 What Statistical Analyses for Interpreting Toxicity Data Are Relevant to State WQS?

The statistical methods to analyze the promulgated WET testing methods, as well as the sediment and bioaccumulation methods, are part of the testing methods themselves and should be strictly followed. One problem states have had with analyzing data is related to not following the correct statistical procedures described in the methods. The state's DQOs and QA/QC procedures should clearly define adequate statistical and other implementation procedures to ensure that all parties are aware of the minimum data set and statistical analysis requirements.

Table 6-1. Hierarchy of toxicological approaches and levels for evaluation of aquatic life use attainment

Level of info[a]	Technical components	Spatial/temporal coverage	Data quality[b]	WBS codes[c]
1	Any one of the following: • Acute or chronic WET • Acute ambient • Acute sediment	1-2 WET tests/yr or 1 ambient or sediment sample tested in a segment or site	Unknown/low; minimal replication used; laboratory quality or expertise unknown	510, 520, 530, 550
2	Any of the following: • Acute or chronic ambient • Acute sediment • Acute and chronic WET for effluent-dominated system	3-4 WET tests/yr or 2 ambient or sediment samples tested in a segment or site at different times	Low/moderate; little replication used within a site; laboratory quality or expertise unknown or low	510, 520, 530, 540, 550
3	Any of the following: • Acute and chronic WET for effluent-dominated system • Chronic ambient or acute or chronic sediment	Monthly WET tests or total of three tests based on samples collected in a segment at three different times	Moderate/high; replication used; trained personnel and good laboratory quality	510, 520, 540, 550
4	Both of the following: • Acute and chronic ambient and • Acute or chronic sediment	≥ 4 tests in total based on samples collected in a segment at four different times including low-flow conditions	High; replication used; trained personnel and good laboratory quality	530, 540, 550

[a] Level of information refers to rigor of toxicity testing, where 1 = lowest and 4 = highest.

[b] Refers to ability of the toxicity testing endpoints to detect impairment or to differentiate along a gradient of environmental conditions.

[c] WBS Assessment Type Codes from Table 1-1.

The procedures should describe how the state uses trend analysis for toxicity monitoring or requirements.

6.3.2 How Does the State Make Attainment/Impairment Decisions in the Absence of a "Perfect" Data Set?

State assessment and listing methodologies should describe the state's efforts to make water quality attainment/impairment decisions in the absence of complete data sets that meet all their data quality requirements. A state should develop procedures for looking for overwhelming evidence of water quality impairment, such as why a single sample (with well-documented QC methods) shows high toxicity.

6.4 References

U.S. Environmental Protection Agency (U.S. EPA). 1988. Guidance for state implementation of water quality standards for CWA section 303(c)(2)(B). Office of Water Regulations and Standards, Criteria and Standards Division, Washington, DC.

U.S. EPA. 1991. Technical support document for water quality-based toxics control. Washington, DC. EPA/505/2-90-001.

U.S. EPA. 1993. Methods for measuring the acute toxicity of effluents and receiving water to freshwater and marine organisms. 4th ed. Office of Research and Development, Washington, DC. EPA/600/4-90/027F.

U.S. EPA. 1994a. Short-term methods for estimating the chronic toxicity of effluents and receiving water to marine and estuarine organisms. 2nd ed. July. Office of Research and Development, Washington, DC. EPA/600/4-91/003.

U.S. EPA. 1994b. Short-term methods for estimating the chronic toxicity of effluents and receiving water to freshwater organisms. 3rd ed. Office of Research and Development, Washington, DC. EPA/600/4-91/002.

U.S. EPA. 1994c. Methods for assessing the toxicity of sediment-associated contaminants with estuarine and marine amphipods. Office of Research and Development, Duluth, MN. EPA-600-R-94-025.

U.S. EPA. 1997. The incidence and severity of sediment contamination in the United States. Volume 1: National sediment quality survey (EPA-823-R-97-006); Volume 2: Data summaries for areas of probable concern (EPA-823-R-97-007); Volume 3: Sediment contaminant point source inventory. (EPA-823-R-97-008). Office of Water, Office of Science and Technology, Washington, DC.

U.S. EPA. 1999. Errata for effluent and receiving water toxicity test manuals: Acute toxicity of effluents and receiving waters to freshwater and marine organisms; Short-term methods for estimating the chronic toxicity of effluents and receiving waters to freshwater organisms; and Short-term methods for estimating the chronic toxicity of effluents and receiving waters to marine and estuarine organisms. January 1999. Office of Research and Development, Duluth, MN. EPA-600/R-98/182.

U.S. EPA. 2000. Methods for measuring the toxicity and bioaccumulation of sediment-associated contaminants with freshwater invertebrates. 2nd ed. Office of Research and Development, Duluth, MN. EPA 600/R-99/064.

U.S. EPA. 2001. Methods for assessing the chronic toxicity of marine and estuarine sediment-associated contaminants with the amphipod *Leptocheirus plumulosus*. March 2001. Office of Water, Office of Science and Technology, Washington, DC. EPA-823-F-01-008.

7. Using Bacteria Criteria as Indicators of Water Quality

Contents

7. Using Bacteria Criteria as Indicators of Water Quality

This chapter addresses the role of bacteria criteria in determining attainment of applicable water quality standards (WQS) and listing impaired waters. The chapter provides a framework for states' assessment and listing methodologies. The information that states should provide about assessment methods includes:

- How are bacteria criteria used within the context of the state's WQS?
- How does the state define and then evaluate the quality of bacteria criteria sets from primary and secondary sources?
- How does the state interpret bacteria criteria to assess WQS attainment or nonattainment?
- How does the state use predictive tools to support attainment decisions?

7.1 How Are Bacteria Criteria Used Within the Context of the State's Water Quality Standards?

Under the Clean Water Act (CWA), states must establish WQS for all waters within their jurisdiction. A WQS defines a use (or uses) for a waterbody and describes the specific water quality criteria to achieve that use. In establishing WQS, states must (1) designate uses consistent with CWA goals, (2) establish water quality criteria to protect the uses, and (3) develop and implement antidegradation policies and procedures. States are also expected to develop implementation procedures for the WQS. These procedures address mechanisms for assessing attainment with the WQS as well as translation of the WQS to an NPDES permit limit.

States generally use bacteria data to develop water quality criteria to protect three designated uses: recreation, shellfish consumption, and public water supply. Responsibility for protecting these designated uses is typically distributed among multiple state and local agencies. Therefore, state water quality agencies frequently rely on interpretation of bacteria data collected and analyzed by other agencies when making designated use support determinations under CWA sections 303(d) and 305(b). For example, state and local public health departments are usually responsible for monitoring to ensure waters are suitable for recreation and shellfish harvesting. Drinking water officials may collect data on the condition of a drinking water supply's source. The state should use the information collected by these different entities in making designated use support decisions.

7.1.1 Recreational Designated Use

Recreational designated uses include swimming, wading, boating, surfing, and other activities in which people come into full or partial contact with surface waters. Waters may be assigned a general recreational use designation. Some states differentiate among types of recreational uses and designate waters for specific subcategories of recreational uses. EPA's *Implementation Guidance for Ambient Water Quality Criteria for Bacteria* provides more detail about options for adopting subcategories of recreational uses (U.S. EPA 2002 - projected).

Based on a risk management approach to protecting recreational uses of surface waters, EPA's recommendation is to use *E. coli* and enterococci as bacteria indicators as described in Ambient Water Quality Criteria for Bacteria - 1986. Table 7-1 summarizes the current section 304(a) criteria recommendations related to bacteria indicators for fresh waters and marine waters. The criteria values are based on levels of risk correlating to no more than 8 cases of acute gastrointestinal illness per 1,000 swimmers for fresh waters and no more than 19 illnesses per 1,000 swimmers for marine waters (U.S. EPA 1986). States may exercise risk management discretion and adopt criteria based on illness rates up to 19 illnesses per 1,000 swimmers. The *Implementation Guidance for Ambient Water Quality Criteria for Bacteria* provides calculations of criteria based on different risk levels.

Many states still use the pre-1986 standard for fecal coliform as the numeric criterion to protect recreational uses. EPA recommends state transition to the *E. coli* and enterococci criteria because these bacteria indicators correlate more strongly to gastrointestinal problems than does the fecal coliform indicator. Recent amendments to the CWA have placed additional requirements on states with coastal and Great Lakes waters to adopt water quality criteria consistent with EPA recommendations for recreational waters to protect beaches specifically. EPA encourages states to adopt these criteria to protect all recreational waters.

7.1.2 Shellfish Consumption Designated Use

Fish and shellfish consumption is a beneficial use protected under the Clean Water Act. Waters where shellfish may be harvested for human consumption are protected by numeric water quality criteria aimed at preventing public health risks associated with bacteria contamination. The numeric human health criteria for water column concentrations of bacteria indicators that are outlined in Quality Criteria for Water (U.S. EPA 1977) may be used as a basis for determining impairment to shellfish waters (see Table 7-1). These criteria are consistent with those used by the National Shellfish Sanitation Program.

7.1.3 Public Water Supply Designated Use

EPA and states protect waters used as drinking water supplies under both the Safe Drinking Water Act (SWDA) and the CWA. Under the SDWA, EPA develops National Primary Drinking Water Regulations. Some of these address source water quality, although most address the quality of treated drinking water. The SDWA also addresses the protection of source waters through state planning activities including the Source Water Protection Plans and Wellhead Protection Plans. Under the CWA, states protect waters designated as public water supplies by adopting criteria sufficient to protect the use.

Table 7-1. Current water quality criteria for bacteria indicators

Designated use	Bacteria indicators evaluated	Criteria
Primary contact recreation	*E. coli*[a]	Freshwater geometric mean: not to exceed 126 CFU per 100 mL, based on no fewer than five samples equally spaced over a 30-day period. Freshwater single-sample maximum: no sample should exceed a one-sided CL calculated using 235 CFU/100 mL (designated bathing beach) 75% CL; 298 CFU/100 mL (moderate use for bathing) 82% CL; 406 CFU/100 mL (light use for bathing) 90% CL; 576 CFU/100 mL (infrequent use for bathing) 95% CL; based on a site-specific log standard deviation, or if site data are insufficient to establish a log standard deviation, then using 0.4 as the log standard.
	Enterococci[a]	Freshwater geometric mean: geometric mean not to exceed 33 CFU per 100 mL, based on no fewer than five samples equally spaced over a 30-day period. Freshwater-single sample-maximum: no sample should exceed a one-sided CL calculated using 61 CFU/100 mL (designated bathing beach) 75% CL; 89 CFU/100 mL (moderate use for bathing) 82% CL; 108 CFU/100 mL (light use for bathing) 90% CL; 151 CFU/100 mL (infrequent use for bathing) 95% CL; based on a site-specific log standard deviation, or if site data are insufficient to establish a log standard deviation, then using 0.4 as the log standard. Marine geometric mean: geometric mean not to exceed 35 CFU per 100 mL, based on no fewer than five samples equally spaced over a 30-day period. Marine single-sample maximum: no sample should exceed a one-sided CL calculated using 104 CFU/100 mL (designated bathing beach) 75% CL; 158 CFU/100 mL. (moderate use for bathing) 82% CL; 276 CFU/100 mL (light use for bathing) 90% CL; 500 CFU/100 mL (infrequent use for bathing) 95% CL; based on a site-specific log standard deviation, or if site data are insufficient to establish a log standard deviation, then using 0.7 as the log standard.
Secondary contact recreation		Adopt criteria commensurate with anticipated use, not to exceed five times the geometric mean values used for primary contact recreation.[b]
Shellfish harvesting	Total coliform[c]	Geometric mean of 70 most probable number (MPN) per 100 mL, with not more than 10% of the samples taken during any 30-day period exceeding 230 MPN per 100 mL.
	Fecal coliform[c]	Median concentration should not exceed 14 MPN per 100 mL, with not more than 10% of the samples taken during any 30-day period exceeding 43 MPN per 100 mL.
Public drinking water sources		Ambient criteria are under development.

[a] Source: U.S. EPA 1987.
[b] Source: U.S. EPA 2002.
[c] Source: U.S. EPA 1977.

Standards applicable to treated drinking water are outlined in 40 CFR 141.71(a)(1). States may wish to adopt these values into their WQS to protect waters designated as public water supplies. EPA is also working on standards that apply to drinking water sources.

The presence of bacteria in treated drinking water, on the basis of fecal indicators or other bacteria indicators, shows that the water may be unsafe for consumption. Thus, EPA has set the maximum contaminant level goal at zero for *Cryptosporidium* and *Giardia lamblia*, total coliforms, and viruses for treated water distributed by public drinking water systems (40 CFR 141.74). Water quality criteria have not been developed for these bacteria indicators in surface waters used as drinking water supply. Most protozoa, viruses, and bacteria are inactivated by chlorine or other disinfectants used during the treatment process, although some human bacteria are more resistant to disinfection than others. All disinfection and filtration technologies are designed to remove a portion, but not all, of bacteria contamination from the influent. Therefore, higher levels of bacteria in the source water potentially translate into contamination levels in the treated water and thus pose a greater public risk.

7.2 What Are the Data Quality and Documentation Requirements for Bacteria Criteria From Primary and Secondary Data Sources?

A state's assessment methodology should document data quality requirements so that all interested parties may contribute relevant data. States should include this information in their standard operating procedures (SOPs) or Quality Assurance Project Plans (QAPPs). The assessment methodology may reference these other documents instead of reiterating the state's requirements.

7.2.1 What Are the State's Requirements for Sample Collection and Analytical Methods And/or Performance Criteria for Analytical Methods?

Adherence to specific procedures for sampling is recommended for a successful and effective water quality monitoring program. Collection, preservation, and storage of water samples are critical to the results of analyses for bacteria indicators. Detailed guidance on sample collection and analytical techniques is available at *http://www.epa.gov/microbes*.

7.2.2 What Is the State's Process for Evaluating the Quality of Bacteria Criteria?

Data quality assessment means the scientific and statistical evaluation of data to determine whether data obtained from monitoring operations are of the right type, quality, and quantity to support water quality assessments. Data quality does not exist in a vacuum; one must know in what context a data set is to be used in order to establish a relevant yardstick for judging whether the data set is adequate.

For assessing WQS attainment, EPA recommends a tiered approach. The following steps should be part of the first tier of the data quality review process.

- Were data generated using appropriate sampling and analysis methods?
- Were samples collected under the appropriate conditions for comparison with WQS (e.g., correct time of year or flow conditions)?
- Is there documentation to show that sampling and analysis results were evaluated according to QAPPs or other project requirements?
- Is there quality control information with the data set?
- What are the limitations of the data, and are the data usable with those limitations?
- Do data meet assessment needs?
- Do the metadata accompanying the data set meet agency standards (e.g., determine adequacy and accuracy of geographic documentation in the data set)?

Once the state determines that the data set meets basic documentation requirements, the data set is ready for analysis to support WQS attainment decisions. In some cases, the state may decide additional screening is necessary before the data set is ready to support attainment decisions. For example, the state may want to look for values below the detection limit of the analytical method, because these may influence how the data set is analyzed or incorporated into other data sets. If upon initial analysis of the data, the findings raise suspicions about possible errors in collection or analysis, the state may want to conduct more in-depth analysis of quality assurance/quality control (QA/QC) procedures. This screening could include reviews of QA/QC reports to determine whether the data set meets the agency's QA/QC requirements regarding: documenting measurement system performance (e.g., adequate use of QC samples), handling missing data and nondetects, and deviations from SOPs. Guidance for assessing the quality of available data sets is described in detail in Practical Methods for Data Quality Assessment (EPA/600/R-96/084).

7.2.3 How Does the State Document the Level of Data Available To Support an Attainment/Impairment Decision?

The 305(b) Consistency Workgroup developed a table assigning qualitative levels of information or data quality to different types of chemical data. Several states have since developed similar approaches for rating the quality of data used in WQS assessments. States are encouraged to use an approach similar to that described in Table 4-2 to report on the quality of data supporting attainment/impairment decisions. In addition, they should begin documenting quantitative information about the quality of these decisions.

The data hierarchy described in Table 4-2 addresses data quality considerations such as sample collection and analytical techniques, spatial and temporal representativeness, and QA procedures. The user rates the data set on the basis of the rigor of the information, where 1 is the lowest and 4 is the highest. In general under this approach, Level 1 information alone is not

sufficient for an attainment decision; however, even a short period of record can indicate impairment in cases of gross exceedances of criteria.

States should supplement the data descriptions in Table 4-2 with more quantitative descriptions of the confidence and power of their attainment/impairment decisions. This documentation clearly illustrates to decision makers and the public the impact of small data sets on uncertainty in the water quality decision. Quantitative documentation of the uncertainty is expressed in statistical terms of the error rates, both Type I decision error, or the α-level, and Type II decision error, or the β-level, of the assessment. These decision errors are discussed in detail in Appendices C and D. A Type I error occurs when an attaining waterbody is erroneously judged to be impaired, and a Type II error occurs when an impaired waterbody is erroneously judged to be attaining. EPA encourages states to collect sufficient numbers of samples to balance both types of error at reasonable levels.

To summarize, for attainment decisions based on chemical data, States should document:

- Level of information based on Table 4-2 or state-developed table or approach
- Sample size, range of concentrations, mean, median, and standard deviation
- Level of statistical confidence (Type I decision error and Type II error) and width of the confidence interval.

7.3 How Does the State Interpret Bacteria Criteria To Assess WQS Attainment/Impairment?

Once the state has assembled data that meet its data quality requirements, it analyzes the data in order to compare them with applicable state standards. The state's consolidated assessment and listing methodology should describe its process for interpreting data to make WQS attainment decisions. This section focuses on data analysis. The next section addresses the use of predictive tools. Chapter 3 deals with integration of bacteria indicator and other data to assess individual designated uses, including recreation, public water supply, and shellfish consumption. It is important that the State's methodology describe all of the approaches it uses to assess data and information when it makes WQS attainment decisions.

7.3.1 How Does the State Process the Data Set?

Water quality measurements from natural systems occur in natural patterns that can be considered a distribution of values. When the data values fall in a systematic pattern around the mean and then taper off rapidly to the tails, it is called a normal distribution or bell-shaped curve. However, in some instances, a log-normal distribution, which has a more skewed (lopsided) shape than normal, occurs. The log-normal is bounded by zero and has a fatter tail than the normal. When sampled data take the shape of a log-normal distribution, it is common practice to transform data to achieve approximate normality prior to conducting statistical tests. In most instances the data values for *E. coli* and enterococci, for example, will be log-normally

distributed, hence the state should transform the data and apply the geometric mean. Appendix C provides additional information on analysis of log-normal data sets.

7.3.2 How Does the State Analyze Bacteria Criteria and Compare the Findings With the Applicable Criteria?

Many states have not yet documented implementation procedures that describe how they will collect and interpret water quality data to assess attainment with applicable WQS in the ambient environment. In lieu of these procedures, states often summarize the available bacteria criteria and compare the results directly with applicable WQS. A majority of states use the recommendations of the 305(b) Consistency Workgroup to assist in making designated use support determinations (U.S. EPA 1997). The CALM document, along with the new implementation guidance for bacteria criteria, is intended to provide a framework for states to build from the 305(b) recommendations, document an assessment and listing methodology, and incorporate it into the state's WQS implementation policies.

Many states are moving forward in documenting how WQS attainment decisions are made. States are developing or revising WQS implementation policies to address the variability and uncertainty associated with monitoring water quality conditions. These implementation policies address:

- Desired confidence levels in the data used to support the decision
- Appropriate size of confidence intervals to control the precision of the data set
- Minimum sample size to control for the potential to conclude that a waterbody is attaining a standard when insufficient data have been collected to credibly detect nonattainment.

All three factors should be included in the state's methodology for interpreting WQS. These factors, in combination, allow the state to reflect the sensitivity of the statistical analysis, balance decision error, and control the precision of WQS attainment decisions. These procedures should be referenced in the state, territory or authorized tribe's approved WQS or in other implementing regulations or policies and procedures documents such as the continuous planning process or consolidated assessment and listing methodology.

Equally important is ensuring that sufficient resources are available for monitoring activities needed to collect data that meet the state's decision making needs. These resources may come from a variety of sources, including state water quality agency budgets, EPA grants, and other monitoring partners in the public, private, academic, and volunteer sectors. Until sufficient resources are in place, states should develop contingencies for how attainment decisions will be made when the data set does not meet these objectives. Some states target these waters for additional monitoring in the short term (see text box).

A state has developed implementation procedures for remote waters that are not designated as public beaches. The procedures use the single-sample maximum as a trigger for collecting five samples within a 30-day period. If the routine monitoring finds an exceedance of a single-sample maximum, then the state collects additional samples to calculate the geometric mean. Then the state uses the geometric mean to make an attainment/nonattainment decision (i.e., both the geometric mean and the single-sample maximum need to exceed the state standard for the waterbody to be identified as impaired under 305(b) and 303(d)).

The 1997 guidelines for preparing 305(b) reports include recommendations for using bacteria criteria in making decisions about attainment of primary contact recreation water quality criteria. The decision rules for interpreting bacteria criteria that were recommended in these guidelines are summarized in Table 7-2. These rules have been modified for presentation in this table by combining the partially supporting and not supporting categories under the impaired category.

The 305(b) guidelines do not contain similar decision rules for assessing bacteria criteria to determine attainment with the shellfish harvesting use or public water supply use. The 305(b) recommendations for assessing attainment with these uses focus on using advisory or closure notices to support attainment decisions. This procedure is discussed in more detail in Chapter 3.

Table 7-2. Using bacteria criteria to support WQS attainment decisions for primary contact recreation

Bacteria indicators	**Attaining WQS**	**Impaired**
Enterococci	Geometric mean and the single-sample maximum are met	Geometric mean is exceeded or single-sample maximum is exceeded during recreational season
E. coli	Geometric mean and the single-sample maximum are met	Geometric mean is exceeded or single-sample maximum is exceeded during recreational season
Fecal coliform	Geometric mean is met and no more than 10% of samples exceed single-sample maximum	Geometric mean is exceeded or more than 10% of samples exceed single-sample maximum

Other environmental data such as salinity, temperature, turbidity, and rainfall might provide further insights into interpretation of the data obtained; however, they do not serve as independent indicators of possible health risks. For example, if the salinity of the water off an ocean beach is lower than normal, perhaps additional flow from a point-source discharge or nonpoint runoff was present. Abnormally cold water temperatures could stress microorganisms collected in the samples, and procedures to recover stressed microorganisms might not accurately measure bacteria indicator density. High turbidity often coincides with a higher density of microbiological organisms. Drinking water treatment plants use information on seasonal fluctuations of turbidity to adjust treatment for optimal removal of microbial contaminants.

The role of rainfall data in predicting bacteria indicator density is discussed in the next section.

7.4 How Does the State Use Predictive Tools To Support WQS Attainment Decisions?

A variety of predictive tools can be used to assess attainment and evaluate the need for beach or shellfish harvesting closures, advisories, and warnings. This section contains descriptions of several such tools, including their attributes, limitations, input data requirements, and availability.

Although most states do not use precautionary advisories or closures when making WQS attainment/impairment decisions, they may use the results of calibrated predictive tools to support WQS attainment/impairment decisions. The assessment and listing methodology should describe how the State uses these tools.

7.4.1 Predictive Models

The tools currently in use by local agencies vary in their complexity and approach to minimizing exposure to pathogens. Simulation of water quality conditions under various scenarios of untreated or partially treated sewage can be used. Comparison of the resulting water quality conditions with the established criteria can serve as the basis for a beach closure. A model was developed for the New York-New Jersey harbor that can predict water quality conditions that result from the bypassing of sewage at preselected locations. Beaches surrounding the discharge location are closed whenever the predicted indicator concentrations exceed the water quality criteria.

Advisories and closures for beaches and shellfish-harvesting areas that are based on water quality modeling are also issued in the states of Virginia, Rhode Island, and Washington. Computer models that predict bacteria indicator concentration by simulating the dominant mixing and transport processes in the receiving water range from simple to very complex. The Virginia Department of Health uses a simple mixing and transport model to predict water quality conditions surrounding wastewater treatment plant outfalls. *Review of Potential Modeling Tools and Approaches to Support the Beach Program* (U.S. EPA 1999) provides a detailed description of these and other tools, and their attributes, limitations, data requirements, and availability. A summary of the capabilities and applicability of these models is included in Table 7-3.

Table 7-3. Evaluation of model capabilities and applicability

Model	Combined PS/NPS	Real time and decision–making	Spills	Application to beach or shellfish closure	Ease of use	Input data required	Calib.	Developing guidelines	Pollutant routing
Rainfall-based	High	High	N/A	High	High	Low	Medium	Medium	N/A
Bypass	Low (PS)	High	High	High	High	High	Low	Medium	High
SMTM	Low (PS)	Medium	Medium	Medium	Medium	Low	Low	N/A	Low
PLUMES	Low (PS)	Low	Medium	Medium	Medium	Low	Low	Low	Low
CORMIX	Low (PS)	Low	Low	Low	Medium	Low	Low	Low	Low
JPEFDC	Medium (NPS/PS)	Low	High	High	Low	High	Medium	Low	High

N/A = Not applicable.

7.4.2 Rainfall-Based Alert Curves

A rainfall-based alert curve is a statistical relationship between the amount of rainfall at representative rainfall gages in the watershed and the observed bacterial indicator concentration at a specific beach area. This relationship is based on simple regression methods and the frequency of exceedance of simultaneous and representative observations of bacterial indicator concentrations and rainfall events. Bacteria criteria supporting the development of rainfall-based alert curves are generated from the water column concentrations obtained from probabilistic or targeted monitoring programs. Although these models do not explicitly account for point and nonpoint sources or fate and transport processes, they rely on a direct statistical relationship and provide simple, easy-to-use tools with reasonable accuracy.

Rainfall-based alert curves based on regression analysis have been used for preemptive beach closures in Milwaukee, WI; Stamford, CT; Sussex County, DE; and the Boston area. The overall objective of beach closure predictive tools is to minimize the population's exposure to bacteria. The tools currently in use vary in their complexity and approach to minimizing exposure, but are generally simple and reliable. The approach taken by the cities of Milwaukee and Stamford and the Delaware Department of Natural Resources and Environmental Control (DNREC) was regression analysis to relate rainfall to bacteria indicator concentration. Models developed on the basis of this approach are site-specific because they are derived from locally observed water quality and rainfall data as well as beach location/configuration relative to bacteria sources. Rainfall-based preemptive shellfish harvesting closures have been used for a number of years in Massachusetts, Rhode Island, and other states.

Rainfall-based alert curve models establish a statistical relationship between rainfall events and bacterial indicator concentrations. This relationship can then serve as a predictive tool to determine the need for beach advisories or closings based on predicted bacteria indicator concentrations. Several agencies have developed beach operating rules based on analysis of site-specific relationships between rainfall and water quality monitoring data. Delaware (DNREC 1997), Wisconsin (Pape 1998), and Connecticut (Kuntz 1998) have successfully used this approach (U.S. EPA 1999).

Rainfall-based alert curves are developed in three phases: collecting data, analysis of data (linking the rainfall events to bacterial indicators), and developing operating rules for advisories or closings of recreational waters. Although EPA is currently supporting continued efforts in research and development of these techniques, the Agency recommends that state and local beach managers consider developing scientifically based and easy-to-use site-specific decision rules based on the technical approach summarized below.

- Rainfall-based models are site-specific, and their development requires relatively large monitoring data sets of both rainfall and water quality. The overall relationship can be described by a statistical regression/estimation model. Depending on the number of rainfall stations considered and the number of rainfall characteristics (e.g., amount, intensity, duration, lag time), the relationship might require a more complex multiple-regression model. Because of the statistical nature of these models, they cannot distinguish between

point sources and nonpoint sources of bacteria indicators and do not explicitly incorporate advection, transport, or decay processes. Because their use is limited to assisting in the development of decision rules for advisories and closings of recreational waters, these models do not attempt to provide the spatial and vertical distribution of bacteria indicators.

- Frequency of exceedance analysis is another rainfall-based method that can be used to develop rainfall-based alert curves. An exceedance is defined as any time the observed bacteria indicator concentration exceeds the action level, such as the state water quality standard, specified by a responsible agency. The objective of this method is to determine the minimum amount of rainfall that causes the bacteria indicator concentration to exceed the action level. This amount can be determined by dividing cumulative rainfall amounts over a period of 24 hours or more into segments that range from no rainfall to an upper limit that is representative of the rainfall record, types of storms, and season. For each rainfall amount category, the observed bacteria indicator concentration or the geometric mean of multiple samples is compared with the action level.

- After a relationship is established between rainfall amounts and bacteria indicator concentrations, developing decision rules for advisories and closings is the next step. An advisory or closing threshold is determined on the basis of the least amount of rainfall that would result in a violation of the action level. This method applies where historical rainfall data and water quality records exist. Decision rules should also be developed to include seasonal variation in rainfall. EPA is currently developing guidance on a number of linear regression techniques that can be used by beach managers to evaluate the need for preemptive advisories or closures.

7.4.3 How Does the State Make Attainment Decisions Using Predictive Models?

EPA recommends using predictive models for making WQS attainment decisions only when the state has collected rainfall data over time and has calibrated and verified the rainfall model. If the model then accurately demonstrates non-attainment under certain conditions and rainfall events, it can be used accurately for making attainment decisions. If the state only uses the model for precautionary closures, the data and model may not be appropriate for making attainment decisions.

7.5 References

American Society for Testing and Materials (ASTM). 1951. Manual on quality control of materials. Special Technical Publication 15-C. Philadelphia: American Society for Testing and Materials.

Cross-Smiecinski A, Stetzenbach LD. 1994. Quality planning for the life science researcher: meeting quality assurance requirements. Boca Raton, FL: CRC Press.

Delaware Department of Natural Resources and Environmental Control (DNREC). 1997. Swimming (primary body contact) water quality attainability for priority watersheds in Sussex County. Delaware Department of Natural Resources and Environmental Control, Dover, DE.

Environment Canada (EC). 2000. Shellfish and water quality. Environment Canada. Dartmouth, Nova Scotia. Accessed March 2001. <http://www.ns.ec.gc.ca/epb/factsheets/sfish_wq.html>.

Gaugush R. 1987. Sampling design for reservoir water quality investigations. Instruction Report E-87-01. Department of the Army, US Army Corps of Engineers, Washington, DC.

Kuntz J. 1998. Non-point sources of bacteria at beaches. City of Stamford Health Department, Stamford, CT.

National Research Center (NRC). 1990a. Monitoring troubled waters: The role of marine environmental monitoring. National Research Center. Washington, DC: National Academy Press.

NRC. 1990b. Monitoring southern California's coastal waters. National Research Center. Washington, DC: National Academy Press.

Pape E. 1998. City of Milwaukee, Wisconsin, Health Department. Personal communication.

U.S. Enviromental Protection Agency (U.S. EPA). 1977. Quality criteria for water. Washington, DC.

U.S. EPA. 1987. Ambient water quality criteria for bacteria - 1986. Office of Research and Development, Microbiology and Toxicology Division, and Office of Water Regulations and Standards, Criteria and Standards Division, Washington, DC.

U.S. EPA. 1997. Guidelines for preparation of the comprehensive state water quality assessments (305(b) reports) and electronic updates. Assessment and Watershed Protection Division, Office of Wetlands, Oceans, and Watersheds, Office of Water. Washington, DC.

U.S. EPA. 1998. The EPA quality manual for environmental programs. EPA Manual 5360. Office of Research and Development, Washington, DC.

U.S. EPA. 1999. Review of potential modeling tools and approaches to support the BEACH program. Office of Science and Technology, Washington, DC. EPA 823-R-99-002.

U.S. EPA. 2000a. Guidance for the data quality objectives process. Office of Research and Development, Washington, DC. EPA 600-R-96-055.
U.S. EPA. 2000b. Guidance: Use of fish and shellfish advisories and classifications in 303(d) and 305(b) listing decisions. Office of Water, Washington, DC. WQSP-00-03.

U.S. EPA. 2002 (projected date). Implementation guidance for ambient water quality criteria for bacteria. Office of Water, Washington, DC.

8. Using Habitat Data as Indicators of Water Quality

Contents

8. Using Habitat Data as Indicators of Water Quality

In the broad sense, physical habitat includes all those structural attributes that influence or provide sustenance to organisms within the waterbody. Physical habitat varies naturally, as do biological characteristics; thus expectations differ even in the absence of anthropogenic disturbance. For example, within a given physiographic region, stream drainage area and overall stream gradient are likely to be strong natural determinants of many aspects of stream habitat, because of their influence on discharge, flood stage, and stream power.

Although habitat quality is recognized as a limiting factor for fisheries and aquatic biota, there is not a consensus on what specific components of physical habitat structure should be assessed. For listing and delisting purposes, the habitat information that is directly linked to pollutants, such as sediments, temperature, and nutrients, will be fundamental. However, identification of impairment related to other physical habitat degradation may be equally important in identifying cost-effective remediation measures and shaping future environmental statutes and regulations.

In developing physical habitat monitoring programs, states should consider several interrelated objectives. Although this guidance is primarily focused onWQS attainment/impairment decisions, physical habitat monitoring programs should address (1) the characterization of physical habitat conditions in reference waterbodies, (2) refinement of aquatic life use classifications in water quality standards (WQS), (3) development and refinement of WQS for specific attributes of physical habitat, and (4) refinement of associated habitat assessment methods.

8.1 How Are Habitat Data Used Within the Context of the States' Water Quality Standards?

Physical habitat assessments are an important part of aquatic life use support determinations for all waterbody types because they (1) facilitate the interpretation of biological data, (2) provide information on nonchemical stressors, and (3) lead to informed decisions regarding problem identification and restoration. For habitat monitoring conducted to assess attainment of the physical, chemical, and biological integrity goals of the Clean Water Act (CWA), the primary purpose is to determine beneficial use support status for WQS, rather than to quantify habitat available for a specific species. In this example from streams, aquatic life use support determination based on habitat assessment data may be categorized as follows:

- *Attaining the WQS and no use is threatened:* 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 (i.e., minimal grazing or disruptive pressure).

- *Attaining some of the designated uses, no use is threatened:* Modification of habitat slight to moderate, usually due to road crossing, limited riparian zones because of encroaching land use patterns, and some watershed erosion. Channel modification slight to moderate.

- *Impaired or threatened for one or more designated uses:* Moderate to severe habitat alteration by channelization and dredging activities, removal of riparian vegetation, bank failure, heavy watershed erosion, or alteration of flow regime.

Most habitat assessment procedures are designed for streams and rivers; however, guidance for habitat assessment of estuaries, lakes, and wetlands is becoming more readily available (U.S. EPA 2000, 1998, 1997, 1994a, 1994b, Adamus and Brandt 1990). Biological assessment of estuaries and lakes should involve evaluating the habitat as well as the biota of the waterbody. In this type of large waterbody assessment, physical habitat not only is composed of solid structures that serve as shelter but also includes chemical, flow, and hydrography components of the waterbody. Monitoring of wetlands has mainly been done through bioassessment; however, the Hydrogeomorphic (HGM) Approach, developed by the Corps of Engineers, uses geographically calibrated wetland morphology, hydrology, and hydrodynamics to assess wetland conditions (Smith et al. 1995, Brinson 1993, Brinson et al. 1995). Additionally, a series of documents recently published by EPA includes a variety of methods used by states to assess wetland condition (U.S. EPA 2002a-i). Because the methods for habitat assessment of streams and rivers are more established, and data produced using these methods are frequently used in determination of attainment, **this chapter addresses the use of habitat data from rivers and streams**. The aforementioned documents should be consulted for more detailed descriptions of the evaluation of habitat in estuaries, lakes, and wetlands.

8.2 What Habitat Indicators Does the State Use To Evaluate Habitat Quality?

Regardless of the method used for assessing stream habitat, the same components of streams are generally investigated. These characteristics include instream habitat, morphology, and riparian characteristics, each of which includes numerous parameters that are selectively measured depending on the method. This section describes these stream characteristics with their parameters. This is not meant to be an in-depth discussion, but rather an overview and introduction.

EPA identified seven general physical habitat attributes important in influencing stream ecology, each of which is naturally variable and also directly or indirectly influenced by anthropogenic activities (U.S. EPA 1993). These attributes are listed below and discussed in the following paragraphs.

- Stream Size - Channel Dimensions
- Channel Gradient
- Channel Substrate Size and Type
- Habitat Complexity and Cover
- Riparian Vegetation Cover and Structure
- Anthropogenic Alterations
- Channel-Riparian Interactions

8.2.1 Stream Size

Stream size is the main determinant of the quantity of habitat available for aquatic organisms. Natural variation and anthropogenic factors influence channel dimensions, flow patterns, and flooding, all of which affect the quantity and quality of habitat. The physical appearance of a stream is a result of channel adjustments according to the magnitude of flow, erosional debris, and basin relief, along with the history of erosion and sediment deposition (Rosgen 1996). Stream channels need to maintain a channel geometry that provides for water and sediment transport at an equilibrium state that results in a relatively stable channel (Bauer and Ralph 1997). Removal of riparian vegetation reduces the structural stability of the stream channel, with negative impacts to fish productivity (Platts 1990, Platts and Nelson 1989). Frequency of pools and riffles, sinuosity, sediment deposition, stream bank condition, and other channel characteristics are naturally variable in streams, but are also influenced by changes in erosional patterns and flow regimes that are due to human activities.

8.2.2 Channel Gradient

Channel gradient is a major influence of river channel morphology and of associated sediment, hydraulic, and biological functions within a stream network (Rosgen 1996). Stream bed and overall habitat stability are influenced by the stream gradient. The impacts of upstream erosion, sediment introduction, and other anthropogenic influences on stream habitat can be evaluated in detail when gradient interactions with sediment transport are understood.

8.2.3 Channel Substrate

Channel substrate is one of the most important determinants of habitat character for fish and macroinvertebrates in streams. Along with bedform, substrate influences the hydraulic roughness and consequently the range of water velocities in the channel. It also influences the size range of interstices that provide living space and cover for macroinvertebrates, sculpins, and other benthic organisms. Because increased erosion and sediment deposition are common to streams in human-influenced areas, sediment characteristics are often sensitive indicators of the effects of human activities on streams.

Fine sediments in streams either move in suspension in the streamflow or are bounced along the bottom (bedload). The size of the particle and the amount of energy in the stream determine which mode of transport will occur within a stream reach. Large amounts of easily transported bedload will fill in pools, form bars on stream bends, and surround gravel and cobble substrates, resulting in altered channel dimensions and flow patterns, as well as reduced habitat and spawning areas for macroinvertebrates and fish (U.S. EPA 1991).

8.2.4 Habitat Complexity and Cover for Aquatic Fauna

Habitat complexity and cover influence the structure and composition of benthic, fish, and periphyton assemblages in streams (Cummins 1974, Platts et al. 1983). The habitat in which biota reside includes natural structures in the stream, such as cobble (riffles), large rocks, fallen

trees, logs and branches, and undercut banks, available as refugia, feeding, or sites for spawning and nursery functions of aquatic macrofauna. A wide variety and/or abundance of submerged structures in the stream provides macroinvertebrates and fish with a large number of niches, thus increasing habitat diversity. As variety and abundance of cover decreases, habitat structure becomes monotonous, diversity decreases, and the potential for recovery following disturbance decreases. Riffles and runs are critical for maintaining a variety and abundance of insects in most high-gradient streams and serving as spawning and feeding refugia for certain fish. The extent and quality of the riffle is an important factor in the support of a healthy biological condition in high-gradient streams. Riffles and runs offer a diversity of habitat through variety of particle size and, in many small high-gradient streams, will provide the most stable habitat.

Snags and submerged logs are among the most productive habitat structures for macroinvertebrate colonization and fish refugia in low-gradient streams (Benke 1984). Large organic debris is especially important in mountainous regions of the Nation. The amount of large woody debris in streams is related to salmonid abundance and distribution. It also aids in reducing channel erosion and buffering sediment inputs by providing sediment storage in headwater streams (Bauer and Ralph 1997).

Fish abundance is related to the diversity of habitats and number and quality of instream pools in stream environments (Kozel and Hubert 1989, Moore and Gregory 1989). Pool filling and destabilization as a result of sedimentation of the substrate can alter habitat structure. Changes in habitat diversity are often associated with adverse impacts to key salmonid-rearing habitats or pools. Pool quality is largely a function of the amount of cover available in slow-velocity waters (Bauer and Burton 1993).

8.2.5 Riparian Vegetation

Riparian vegetation serves as a buffer to pollutants entering a stream from runoff, controls erosion, and provides woody debris that acts as habitat and dissipates energy in the stream (USDI 1995). Additionally, riparian canopy cover over a stream is important in moderating stream temperatures through shading and as an indicator of conditions that control bank stability and the potential for inputs of coarse and fine particulate organic material (Youngblood et al. 1985). Organic inputs from riparian vegetation become food for stream organisms and structure to create and maintain complex channel habitat. A relatively undisturbed riparian zone supports a robust stream system; narrow riparian zones occur when roads, parking lots, fields, lawns, bare soil, rocks, or buildings are near the stream bank. Residential developments, urban centers, golf courses, and rangeland are the common causes of anthropogenic degradation of the riparian zone. However, the presence of “old field” (i.e., a previously developed field not currently in use), paths, and walkways in an otherwise undisturbed riparian zone may be judged to be inconsequential to altering the riparian zone.

8.2.6 Anthropogenic Alterations and Disturbances

Land use, buildings, and other evidence of human activities in the stream channel and its riparian zone may, in themselves, serve as habitat quality indicators; they may also serve as diagnostic

indicators of anthropogenic stress. In channels, disturbances include channel revetment, pipes, straightening, bridges, culverts, and accumulation of trash. Near-channel riparian disturbances include buildings, lawns, roads, pastures, orchards, and row crops.

8.3 How Does the State Use Different Types of Habitat Assessment Methods to Assess and Document Data Quality?

As with biological sampling, standardization of habitat assessment methods is important to ensuring data validity and reliability (Barbour et al. 2000). Standard operating procedures (SOPs) that describe in detail the criteria for assessing physical habitat, and a QA/QC plan that identifies accountability throughout the assessment process, should be developed to minimize bias, reduce error, and maintain data integrity. Proper training is also important to ensuring that habitat results, especially those developed through visual assessments, are reliable (Hannaford et al. 1997). The four levels of rigor for habitat assessments (Table 8-1) are associated with different types of documentation because of their purposes in assessments. Documentation follows a similar pattern to that used in the four levels of bioassessment, with the lower levels involving little documentation or training and few QA/QC procedures, and higher levels involving more training, detailed SOPs, and complete QA/QC plans.

The following discussion is limited to those physical habitat assessment methods that are commonly used by State agencies in streams and wadeable rivers. Four levels of rigor have been identified for habitat assessment and are summarized in Table 8-1. The level of effort required to assess and document physical habitat degradation will vary, depending on the extent of degradation. The discussions in each assessment level describe the level of data quality and rigor provided and the relationship between data quality and WQS attainment/impairment decisions. A discussion of the data quality and the use of habitat data in making WQS attainment/impairment decisions should be clearly documented in the state, territory or authorized tribe's assessment and listing methodology.

8.3.1 Level 1: Qualitative Visual Characterization of Habitat

Level 1 represents the most qualitative habitat assessment methods, in which physical habitat features are visually characterized and assessment of quality or condition is more of a cursory examination. These more subjective methods described in Level 1 are best suited for screening and site reconnaissance. This level of effort may be adequate to support listing of impaired waters when habitat impairment is grossly apparent, but may not be adequate to provide the sensitivity needed to identify subtle impairments and support delisting decisions. Photodocumentation is especially important at this level of effort, because of the absence of quantitative measurements and estimates.

8.3.2 Level 2: Visual-Based Habitat Assessment

These methods follow the documented Rapid Bioassessment Protocols (RBPs) (U.S. EPA 1999) or similar techniques developed by the respective state/tribal agency that include a combination of physical structural features of the stream and its flood plain (Ball 1982, Ohio EPA 1987, U.S.

Table 8-1. Hierarchy of habitat assessment approaches for evaluation of aquatic life use attainment

Level of Info[a]	Technical Components	Spatial/Temporal Coverage[b]	Data Quality[c]
1	Visual observation of habitat characteristics; no true assessment; documentation of readily discernible land use characteristics that might alter habitat quality; no reference conditions	Sporadic visits; sites are mostly from road crossings or other easy access	Unknown or low precision and sensitivity; professional scientist (biologist, hydrologist) not required
2	Visual observation of habitat characteristics and simple assessment; use of land use maps for characterizing watershed condition; reference condition pre-established by professional scientist	Limited to annual visits and non-specific to season; generally easy access; limited spatial coverage and/or site-specific studies	Low precision and sensitivity; professional biologist or hydrologist not involved, or only correspondence
3	Visual-based habitat assessment using SOPs; may be supplemented with quantitative measurements of selected parameters; conducted with bioassessment; data on land use compiled and used to supplement assessment; reference condition used as a basis for assessment	Assessment during a single season usually the norm; spatial coverage may be limited or broad and commensurate with biological sampling; assessment may be regional or site-specific	Moderate precision and sensitivity; professional biologist or hydrologist performs survey or provides oversight and training
4	Assessment of habitat based on quantitative measurements of instream parameters, channel morphology, and floodplain characteristics; conducted with bioassessment; data on land use compiled and used to supplement assessment; reference condition used as a basis for assessment	Assessment during 1-2 seasons; spatial coverage usually broad and commensurate with biological sampling; assessment may be regional or site-specific	High precision and sensitivity; professional biologist or hydrologist performs survey and assessment

NOTE: Table is based on use in lotic systems. With some modification, these approaches would apply to other waterbody types.
[a] Level of information refers to rigor of habitat assessment, where 1 = lowest and 4 = highest.
[b] Refers to ability of endpoints to detect impairment or to differentiate along a gradient of environmental conditions.
[c] WBS assessment type codes from Table 1-1.

EPA 1989, Barbour and Stribling 1991, 1994, Rankin 1991, 1995, Raven et al. 1998). The key element to implementing these methods for Level 2 is to adhere to strict protocols and training to reduce subjectivity and investigator bias in the method. *In situ* measurements of temperature, canopy cover, and flow, as well as pebble counts for estimates of substrate typing, may accompany habitat assessment. However, most quantitative measurements are lacking in Level 2. EPA recommends using this level to support listing and delisting decisions when field crews are well trained in using the habitat assessment technique, the judgment criteria are calibrated for the stream classes under study, and periodic quality assurance checks are conducted (U.S. EPA 1999). The level of uncertainty and potential for decision errors may be reduced over time by regionalizing the habitat assessment protocols and criteria to reflect conditions in localized reference waterbodies.

Level 2 in Table 8-1 represents visual assessment methods that score the quality of the habitat feature being evaluated, based on the expertise of trained stream ecologists. In the RBPs, parameters are visually assessed and rated on a numerical scale of 0 to 20 (highest) for each sampling reach (U.S. EPA 1989). The ratings are then totaled and compared to a geographically calibrated reference condition to provide a final habitat ranking. Scores increase as habitat quality increases. To ensure consistency in the evaluation procedure, descriptions of the physical parameters and relative criteria are included in the rating form.

The ability to accurately assess the quality of the physical habitat structure using any visual-based approach depends on several factors (U.S. EPA 1999):

- The parameters selected need to represent the various relevant features of habitat structure.

- A clearly defined continuum of conditions for each parameter must exist and the parameter must be characterized from the optimum for the region or stream type under study to the poorest situation reflecting substantial alteration from anthropogenic activities.

- The judgment criteria for the attributes of each parameter should minimize subjectivity through either quantitative measurements or specific categorical choices.

- The investigators need to be experienced in, or adequately trained for, stream assessments in the region under study.

- Adequate documentation and ongoing training must be maintained to evaluate and correct errors.

In the RBPs, habitat evaluations are made on a combination of instream habitat, channel morphology, bank structural features, and riparian vegetation for a total of 10 parameters that vary according to stream gradient (two categories, high or low gradient).

- Epifaunal substrate/available cover, high and low gradient
- Embeddedness, high gradient
- Pool substrate characterization, low gradient

- Velocity/depth combinations, high gradient
- Pool variability, low gradient
- Sediment deposition, high and low gradient
- Channel flow status, high and low gradient
- Channel alteration, high and low gradient
- Frequency of riffles (or bends), high gradient
- Channel sinuosity, low gradient
- Bank stability, high and low gradient
- Bank vegetative protection, high and low gradient
- Riparian vegetative zone width, high and low gradient

Generally, a single comprehensive assessment is made that incorporates features of the entire sampling reach as well as selected features of the catchment. Additional assessments may be made on neighboring reaches to provide a broader evaluation of habitat quality for the stream ecosystem. The actual habitat assessment process involves rating the 10 parameters as optimal, suboptimal, marginal, or poor based on the criteria included on the Habitat Assessment Field Data Sheets. Some state programs, such as the Florida Department of Environmental Protection (Florida DEP1996) and Mid-Atlantic Coastal Streams Workgroup (MACS) (U.S. EPA 1996), have adapted this approach using somewhat fewer and different parameters.

8.3.3 Level 3: Visual-Based Assessment of Habitat With Quantitative Measurements

Level 3 methods are similar to those described in Level 2 in that they provide a rapid, visual-based habitat assessment approach, designed to describe the overall quality of the physical habitat. However, in addition to the visual-based assessment, Level 3 methods are supplemented with Level 4 quantitative measures of channel dimensions, substrate size and type, habitat characteristics, or riparian features. These quantitative methods are more fully described in Level 4. Of particular importance are measurements of sediment loads and excess water temperatures. As described in Table 8-1, the data produced by Level 3 methods are a more reliable description of the habitat at a site than the data produced by Level 1 or 2 methods.

8.3.4 Level 4: Quantitative Assessment of Habitat

Level 4 contains the most quantitative techniques (Table 8-1) that incorporate measurements of various features of the instream, channel, and bank morphology such as the Environmental Monitoring and Assessment Program (EMAP) of EPA and the National Water Quality Assessment Program (NAWQA) of the U.S. Geological Survey (Meader et al. 1993, U.S. EPA 1994c, 1999). These techniques provide a relatively comprehensive characterization of the physical structure of the stream sampling reach and its surrounding floodplain. Quantitative habitat assessments require detailed measurements of stream components by trained biologists and hydrologists. Because Level 4 methods require a high level of expertise and because measurements are detailed and extensive, the data produced by these methods exhibit high precision and reliability. The components that are measured according to EMAP surface waters physical habitat protocol (U.S. EPA 1999b) include:

- Longitudinal profile
- Large woody debris
- Channel and riparian cross-sections
- Discharge

Assessment of these stream components according to EMAP guidelines requires detailed measurements. For longitudinal profile assessment, aquatic habitat is classified and 11 cross-sectional measurements for soft/small sediment and 11 wetted width cross-sectional measurements are taken. To assess large woody debris, size classes of woody debris are counted. For channel and riparian analyses, within 11 cross-sectional areas, measurements are taken of channel dimensions and channel morphologic characteristics, substrate size and type, habitat complexity, and nearby human disturbances. For discharge calculations, 15-20 measurements of depth and velocity are taken at equally spaced intervals across one appropriately chosen cross-section.

In most quantitative methods, these data are then condensed to stream reach summaries that describe particular aspects of physical habitat. Some of these metrics calculated according to EMAP methods include:

- Channel morphology statistical summaries
- Channel cross-section and bank morphology
- Sinuosity
- Slope
- Residual pool analysis (useful for measuring sediment loadings)
- Substrate size and composition
- Bed substrate stability
- Fish cover
- Large woody debris
- Riparian canopy cover (useful for interpreting excess water temperatures)
- Riparian vegetation structure
- Riparian human disturbances

As with the biological data, reference conditions are established as benchmarks and the habitat metrics for test sites are compared to these reference conditions to assess habitat condition and identify impairments. Level 4 habitat assessments include measurements of sediment loadings and excess temperatures.

8.4 How Does the State Analyze Habitat Data To Determine WQS Attainment?

Habitat quality is a required and integral element of water resource integrity. It is clearly an important limiting factor for aquatic biota. However, the lack of consensus on how to quantify and evaluate habitat has hampered the application of indicators of habitat quality, especially in water resource management. Much of the work on habitat to date has focused on the objective of describing potential fisheries (salmonid) production and limits to this production. This is a narrower objective than describing physical/biological integrity under the CWA.

Any credible and feasible approach for use of habitat indicators should include a clear statement of the use of habitat indicators within the context of the CWA and relevant state laws and if not documented in the WQS, should be documented in the assessment and listing methodology. Habitat may or may not be the sole indicator of WQS attainment in state law, however, how it is used in conjunction with other parts of the appropriate standard should be documented. Application of habitat indicators at an appropriate scale and stratification should be meaningful in comparison to reference conditions, with an emphasis on quantitative measures. Using habitat structure as a direct measure of water resource integrity will improve the linkage to diffuse source activities (Bauer and Ralph 1997). Poff and Ward (1990) address the rationale for using physical habitat as a template for stream biota:

> *In lotic ecosystems, physical habitat structure is of critical importance to the distributions and abundances of organisms. In general, greater spatial heterogeneity at the scale of the organisms results in greater microhabitat and hydraulic diversity and hence in greater biotic diversity.*

Bauer and Ralph (1997) suggest a two-part approach (coarse and fine scale) in the development of habitat indicators. In the first part, habitat indicators should be developed at the coarse ecoregional scale and can serve as default numerical indicators until such time that a finer scale analysis determines they are inappropriate. Habitat quality indicators can be developed from reference condition data in much the same way as biological indicators, and the same limitations will apply to any given approach. The selection of habitat parameters should be evaluated in a manner similar to the selection of biological metrics. The most efficient way to select habitat parameters is to review the existing literature and evaluate existing data for responsiveness of habitat parameters to changes in human influence.

In the second part, indicators should be developed at a finer scale (i.e., from basin to watershed scale). Indicators at a finer scale reduce the problems of resolution associated with larger scale indicators. The two scales of indicators are complementary and should proceed at the same time. Development of indicators at small scales is time-consuming and costly. However, the depth of knowledge that these indicators can provide cannot be replaced by indicators at the larger scale. Because of the importance of stream gradient in categorizing a waterbody (by virtue of its relationship to landform, topography, geological formations, elevation, etc.), the RBPs included different protocols for riffle/run prevalence (high gradient) and glide/pool prevalence (low gradient).

8.4.1 Natural Classification of Waterbodies

Classification helps to evaluate the natural variability in waterbodies and to distinguish natural variability from that of human-induced changes. States should separate the data into different categories so that comparisons and evaluations are made on “like” data. However, too many classification categories can make water resource management cumbersome and difficult to explain to the public. Classification of waterbodies occurs at many scales. First there is the coarsest level, which is merely separating data into waterbody types such as lakes, streams and small rivers, large rivers, wetlands, and estuaries. Next, within any given type, it is usually

necessary to further classify waterbodies into homogeneous groups to characterize how ecosystems differ (among the groups) in ecologically important attributes. It would be highly unlikely that any state would be so homogeneous that the physical habitat condition of a particular waterbody type would be uniform throughout the state.

Recent research has suggested that classification consisting of a combination of landscape and physical stream features is an effective approach (Waite et al. 2000, Hawkins et al. 2000). One of the most common landscape classifications is ecoregions. Ecoregions are areas of relative ecosystem homogeneity (or similar quality) defined by similarity of land form, soil, vegetation, hydrology, and general land use. For example, the physical habitat of streams of a given ecoregion are more similar to one another than they will be to streams in another ecoregion. However, ecoregions are not the only method for classifying freshwater ecosystems. Hawkins et al. (2000) point out that the amount of biotic variation related to landscape features is not large, and augmenting classifications based on local habitat features accounts for substantially more variation than the larger scale environmental features. Some states have used other landscape factors such as elevation and rainfall to classify their waterbodies (Spindler 1996).
Stream classification is usually based on a combination of physical features that are used to categorize streams into similar groups. Waite et al. (2000) found that variables such as stream order, stream gradient, or other physical features strengthened classifications based on combinations of ecoregions and catchment. It is likely that classifications will differ regionally in order to obtain the greatest precision and resolution for water quality management. Classifications are improved if collaborations across jurisdictional boundaries are implemented to enhance ecological resolution.

Stream classification is helpful to evaluate the natural variability in streams and distinguish it from human-induced changes (Barbour and Stribling 1991). In the western United States, classification systems are gaining wide use among land managers and other water resource management agencies (Naiman et al. 1992). The two most commonly used systems are those of Montgomery and Buffington (1993) and Rosgen (1985). Montgomery and Buffington (1993) developed a system as part of the State of Washington's Timber, Fish, and Wildlife agreement, and it classifies streams as sediment source, transport, and response (deposition). Rosgen (1985) developed a classification system based on geomorphic and in-channel characteristics. Features of this system include channel gradient, sinuosity, width/depth ratio, bed material, entrenchment, channel confinement, soil erodibility, and bank stability. Rosgen's system is also used to evaluate channel stability. Rosgen's stream-type classification system has been used widely in the west for over ten years (Naiman et al. 1992).

- Level 1 information—no classification of ecosystems done.

- Level 2 information—classification minimal and limited to individual watersheds or basins. May not recognize stream continuum principles where headwaters differ in function from mainstem. In estuaries and lakes, classification may be portions or embayments; however, habitat structure in large ecosystems not well defined.

- Level 3 information—classification done to recognize geographical or other similar organization. Usually based on landscape features and supplemented with instream or other waterbody characteristics.

- Level 4 information—combination of landscape features and physical habitat structure of waterbody type to provide the best classification scheme for assessment.

8.4.2 How Does the State Establish Reference Conditions?

Reference conditions serve as a benchmark for assessing physical habitat condition and establishing water quality goals. The reference condition can be derived from reference sites or some empirical model of expectations that may include knowledge of historical condition or extrapolation from ecological principles. The norm is to use actual sites that represent best attainable conditions of a waterbody, as is done with biological data. In fact, the same reference conditions are often used for habitat and biological data. The discussion of reference conditions in Chapter 5 also applies to those established for assessing physical habitat.

- Level 1 information (See Table 8-1)—no formal reference condition may exist. Professional opinion may be used to support assessment of quality of site and may be unacceptable for listing and delisting if not supported by scientific evidence.

- Level 2 information—reference conditions may be pre-established by professional and based on known physiography of area. Assessments are based on a percentage of reference condition (usually determined by a composite of regional sites) or percentage of maximum score of assessment method.

- Level 3 information—reference condition is regional and/or oriented toward watershed-scale assessments. Regional reference sites are likely developed for the relevant waterbody type and are the basis for habitat assessment and monitoring of changes.

- Level 4 information—regional reference conditions established for each waterbody class and consist of sites and/or other means of establishing regional expectations for assessment of habitat quality.

8.4.3 Data Analysis

Habitat assessment data are important for measuring the attainment of WQS for the protection of aquatic life. The level of effort and approach to data analysis should be commensurate with the specific objectives of the study and level of sampling effort. The lower levels of effort require limited data assessment and, as such, may yield a greater level of uncertainty surrounding WQS attainment decisions. The higher levels of effort, which support the determination of attainment or nonattainment, the development of numeric criteria for physical attributes associated with habitat, and the refinement of assessment methods, may require more rigorous data analyses. The rapid bioassessment protocols describe data assessment and interpretation applicable to the rapid, visual-based methods (U.S. EPA 1999). Procedures for calculating habitat metrics that are

useful for making comparisons among sites or against a baseline or reference condition are described in U.S. EPA 1999b.

- Level 1 information—no formal habitat assessment endpoint is established. Assessment may be based on only on the best professional judgment of the investigator.

- Level 2 information—a habitat assessment endpoint is established for specific waterbodies, but may not be calibrated to waterbody classes or statewide application. Watershed monitoring should be conducted where regional reference conditions have not been established.

- Level 3 information—a habitat assessment has been developed and calibrated for use throughout the state or region for the various classes of a given waterbody type. Index is usually relevant to both quantitative and visual-based measures, but may or may not be applicable among several states or tribes.

- Level 4 information—quantitative habitat measurements are used to assess the physical habitat structure. Based on regional and geomorphological expectations, various degrees of impairment are determined.

8.5 References

Adamus PR, Brandt K. 1990. Impacts on quality of inland wetlands of the United States: A survey of indicators, techniques, and applications of community level biomonitoring data. U.S. Environmental Protection Agency, Environmental Research Laboratory, Corvallis, OR. EPA/600/3-90/073.

Ball J. 1982. Stream classification guidelines for Wisconsin. Wisconsin Department of Natural Resources Technical Bulletin. Wisconsin Department of Natural Resources, Madison, WI.

Barbour MT, Swietlik WF, Jackson SK, Courtemanch DL, Davies SP, Yoder CO. 2000. Measuring the attainment of biological integrity in the USA: A critical element of ecological integrity. Hydrobiologia 422/423:453-464.

Barbour MT, Stribling JB. 1994. A technique for assessing stream habitat structure. In: Conference proceedings, Riparian ecosystems in the humid U.S.: Functions, values and management. National Association of Conservation Districts, Washington, DC, March 15-18, 1993, Atlanta, GA. pp. 156-178.

Barbour MT, Stribling JB. 1991. Use of habitat assessment in evaluating the biological integrity of stream communities. In: Biological criteria: research and regulation, proceedings of a symposium, 12-13 December 1990, Arlington, VA. U.S. Environmental Protection Agency, Office of Water, Washington, DC. EPA-440/5-91/005.

Bauer SB, Burton TA. 1993. Monitoring protocols to evaluate water quality effects of grazing management on western rangeland streams. U.S. Environmental Protection Agency, Region 10. Seattle, WA. EPA-910/R-93-017.

Bauer SB, Ralph S. 1997. Development of habitat quality indicators within the Clean Water Act. November 1997. U.S. Environmental Protection Agency, Region 10. Seattle, WA.

Benke AC, Van Arsdall TC Jr., Gillespie DM. 1984. Invertebrate productivity in a subtropical blackwater river: The importance of habitat and life history. Ecol Monogr 54(1):25-63.

Brinson MM. 1993. A Hydrogeomorphic Classification for Wetlands. Wetland Research Program Technical Report WRP-DE-4. U.S. Army Engineer Waterways Experiment Station, Vicksburg, MS.

Brinson MM, Hauer FR, Lee LC, Nutter WL, Smith RD, Whigham D. 1995. Guidebook for application of hydrogeomorphic assessments to riverine wetlands (operational draft). Wetland Research Program Technical Report WRP-DE-11. U.S. Army Engineer Waterways Experiment Station, Vicksburg, MS.

Cummins KW. 1974. Structure and function of stream ecosystems. Bioscience 24: 631-641.

Florida Department of Environmental Protection (DEP). 1996. Development of the stream condition index (SCI) for Florida. Florida Department of Environmental Protection, Tallahassee, FL.

Hannaford MJ, Barbour MT, Resh VH. 1997. Training reduces observer variability in visual-based assessments of stream habitat. J N Am Benthol Soc 16(4):853-860.

Hawkins CP, Norris RH, Gerritsen J, Hughes RM, Jackson SK, Johnson RK, Stevenson RJ. 2000. Evaluation of the use of landscape classifications for the prediction of freshwater biota: Synthesis and recommendations. J N Am Benthol Soc 19(3):541-556.

Helgen J. 2001. Methods for evaluating wetland condition: Developing an invertebrate index of biological integrity for wetlands. U.S. Environmental Protection Agency; Office of Water; Washington, DC. EPA 822-R-02-019.

Kozel SJ, Hubert WA. 1989. Factors influencing the abundance of brook trout (*Salvelinus fontinalis*) in forested mountain streams. J Freshw Ecol 5(1):113-122.

Meader MR, Hupp CR, Cuffney TF, Gurtz ME. 1993. Methods for characterizing stream habitat as part of the national water quality assessment program. U.S. Geological Survey, Raleigh, NC. USGS/OFR 93-408.

Montgomery DR, Buffington JM. 1993. Channel classification, prediction of channel response, and assessment of channel condition. Report TFW-SH10-93-002, prepared for the SHAMW committee of the Washington State Timber/Fish/Wildlife Agreement. 84 p.

Moore KMS, Gregory SV. 1989. Geomorphic and riparian influences on the distribution and abundance of salmonids in a Cascade mountain stream. USDA Forest Service Gen Tech Rep PSW-110.

Naiman RJ, Lonzarich DG, Beechie TJ, Ralph SC. 1992. General principles of classification and the assessment of conservation potential in rivers. In: River conservation and management. John Wiley, pp. 93-123.

Ohio Environmental Protection Agency (Ohio EPA). 1987. Biological criteria for the protection of aquatic life: volumes I-III. Columbus, OH: Ohio Environmental Protection Agency.

Platts WS. 1990. Managing fisheries and wildlife on rangeland grazed by livestock: A guidance and reference document for biologists. Nevada Department of Wildlife.

Platts WS, Nelson RL. 1989. Characteristics of riparian plant communities and streambanks with respect to grazing in northeastern Utah. In: Gressell RE, et al., eds. Riparian resource management. U.S. Bureau of Land Management, Billings, MT. pp. 73-81.

Platts WS, et al. 1983. Methods for evaluating streams, riparian and biotic conditions. General Tech Rep INT-138. U.S. Department of Agriculture, U.S. Forest Service, Ogden, UT.

Poff NL, Ward JV. 1990. Physical habitat template of lotic systems: recovery in the context of historical pattern of spatiotemporal heterogeneity. Environ Manage 14(5):629-646.

Rankin ET. 1995. Habitat indices in water resource quality assessments. In: Davis WS, Simon TP, eds. Biological assessment and criteria: Tools for water resource planning and decision making. Boca Raton, FL: Lewis Publishers, pp. 181-208.

Rankin ET. 1991. The use of the qualitative habitat evaluation index for use attainability studies in streams and rivers in Ohio. In: USEPA 1991. Biological criteria: Research and regulation, Office of Water, U.S. Environmental Protection Agency, Washington, DC. EPA/440/5-91-005.

Raven PJ, Holmes NTH, Dawson FH, Fox PJA, Everard M, Fozzard IR, Rouen KJ. 1998. River habitat quality - the physical character of rivers and streams in the UK and Isle of Man. Environment Agency.

Rosgen DL. 1996. Applied river morphology. Wildland Hydrology, Pagosa Springs, CO.

Rosgen DL. 1985. A stream classification system. In: Proceedings of the First North American Riparian Conference, Riparian Ecosystems and their Management: Reconciling conflicting uses. USDA Forest Service, Tucson, AZ. GTR RM-120.

Smith RD, Ammann A, Bartoldus C, Brinson MM. 1995. An approach for assessing wetland functions using hydrogeomorphic classification, reference wetlands, and functional indices. Wetland Research Program Technical Report WRP-DE-9. U.S. Army Engineer Waterways Experiment Station, Vicksburg, MS.

Spindler P. 1996. Using ecoregions for explaining macroinvertebrate community distribution among reference sites in Arizona, 1992. Arizona Department of Environmental Quality, Hydrologic Support and Assessment Section, Flagstaff, AZ.

U.S. Department of the Interior. 1993. Riparian area management, process for assessing proper functioning condition. U.S. Department of the Interior, Bureau of Land Management. TR 1737-9.

U.S. Environmental Protection Agency (U.S. EPA). 1989. Rapid bioassessment protocols for use in streams and rivers: benthic macroinvertebrates and fish. Assessment and Watershed Protection Division, Washington, DC. EPA/440/4-89/001.

U.S. EPA. 1991. Monitoring guidelines to evaluate effects of forestry activities on streams in the Pacific Northwest and Alaska. Region 10, Seattle WA. EPA/910/9-91-001.

U.S. EPA. 1993. Physical habitat. In: Hughes RM, ed. Stream Indicator and Design Workshop. Office of Research and Development, Corvallis, OR. pp. 59-69. EPA/600/R-93/138.

U.S. EPA. 1994a. EMAP surface waters field operations manual for lakes. Environmental Monitoring Systems Laboratory, Las Vegas, NV.

U.S. EPA. 1994b. Environmental Monitoring and Assessment Program: integrated quality assurance project plan for the Surface Waters Resource Group, 1994 activities, Rev. 2.00. Las Vegas, NV. EPA 600/X-91/080.

U.S. EPA. 1994c. Environmental Monitoring and Assessment Program 1994 pilot field operations manual for streams. Office of Research and Development, Cincinnati, OH. EPA/620/R-94/004.

U.S. EPA. 1997. Habitat assessment. In: Baker JR, Peck DV, Sutton DW, eds. Environmental Monitoring and Assessment Program - Surface Waters: Field Operations Manual for Lakes. Washington, DC. EPA 620-R-97-001.

U.S. EPA. 1998. Lake and reservoir bioassessment and biocriteria. Office of Wetlands, Oceans, and Watersheds, Office of Science and Technology, Office of Water, Washington, DC. EPA 841-B-98-007.

U.S. EPA. 1999. Rapid bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates and fish. 2nd ed. Office of Water, Washington, DC. EPA 841-B-99-002.

U.S. EPA. 1999b. Quantifying physical habitat in wadeable streams. Kaufmann PR, Levine P, Robison EG, Seeliger C, Peck DV. Office of Research and Development, Washington, DC. EPA/620/R-99/003.

U.S. EPA. 2000. Estuarine and coastal marine waters: Bioassessment and biocriteria technical guidance. Office of Water, Office of Science and Technology, Washington, DC. EPA 822-B-00-024.

U.S. EPA. 2002a. Methods for evaluating wetland condition: Introduction to wetland biological assessment. Office of Water; Washington, DC. EPA 822-R-02-014.

U.S. EPA. 2002a. Methods for evaluating wetland condition: Introduction to wetland biological assessment. Office of Water; Washington, DC. EPA 822-R-02-014.

U.S. EPA. 2002b. Methods for evaluating wetland condition: Study design for monitoring wetlands. Office of Water; Washington, DC. EPA 822-R-02-015.

U.S. EPA. 2002c. Methods for evaluating wetland condition: Developing metrics and indexes of biological integrity. Office of Water; Washington, DC. EPA 822-R-02-016.

U.S. EPA. 2002d. Methods for evaluating wetland condition: Wetlands classification. Office of Water; Washington, DC. EPA 822-R-02-017.

U.S. EPA. 2002e. Methods for evaluating wetland condition: Developing an invertebrate index of biological integrity for wetlands. Office of Water; Washington, DC. EPA 822-R-02-019.

U.S. EPA. 2002f. Methods for evaluating wetland condition: Using vegetation to assess environmental conditions in wetlands. Office of Water; Washington, DC. EPA 822-R-02-020.

U.S. EPA. 2002g. Methods for evaluating wetland condition: Using algae to assess environmental conditions in wetlands. Office of Water; Washington, DC. EPA 822-R-02-021.

U.S. EPA. 2002h. Methods for evaluating wetland condition: Using amphibians in bioassessments of wetlands. Office of Water; Washington, DC. EPA 822-R-02-022.

U.S. EPA. 2002i. Methods for evaluating wetland condition: Biological assessment for birds. Office of Water; Washington, DC. EPA 822-R-02-023.

Waite IR, Herlihy AT, Larsen DP, Klemm DJ. 2000. Comparing strengths of geographic and nongeographic classifications of stream benthic macroinvertebrates in the Mid-Atlantic Highlands, USA. J N Am Benth Soc 19(3):429-441.

Youngblood AP, Padgett WG, Winward AH. 1985. Riparian community type classification of Eastern Idaho - Western Wyoming. U.S. Department of Agriculture. R4-Ecol-85-01.

Part B — Integrated Monitoring Design for Comprehensive Assessment and Identification of Impaired Waters

Contents

10. Selecting Metrics or Indicators of Water Quality Standards Attainment

Contents

10. Selecting Metrics or Indicators of Water Quality Standards Attainment

This chapter provides recommendations for selection of water quality indicators to serve as measures of water quality standards (WQS) attainment status. These recommendations are based on the report of the Intergovernmental Task Force on Monitoring Water Quality (ITFM 1995). They serve as a starting point for states as they tailor selection of indicators according to their WQS and data quality objectives (DQOs). This chapter also presents considerations for identifying additional or supplemental indicators that could be included in followup or site-specific monitoring.

The first activity a state may undertake in designing a water quality monitoring framework is identifying the appropriate indicators and their endpoints for making attainment/impairment decisions. The state, territory, or authorized tribe's WQS drive this selection process. The state should have a mechanism for interpreting data on water quality indicators within the context of its standards, including designated uses, narrative or numeric criteria, and antidegradation policies. This mechanism may be referenced in the state, territory, or authorized tribe's approved WQS or alternatively in other implementing regulations or policies or procedures documents such as the continuous planning process or consolidated assessment and listing methodology. Other factors that influence a state's selection of indicators and are related to the sampling effort include the cost of collecting and analyzing samples, the variability of the indicator in the environment, the level of precision desired by decision makers, and the sampling frequency required to meet the DQOs (U.S. EPA 1991).

Indicators could include chemicals, biological indices, fish tissue action levels, risk assessment levels, and other measures used to assess attainment with WQS. The monitoring design framework recognizes that selection of indicators is part of an iterative process that also includes establishing appropriate monitoring sites or locations. Key elements of the state's assessment methodology are identification of core or first-tier indicators and a process for developing supplemental indicators.

Limited resources will affect actions and decisions for many water quality monitoring programs. Optimal use of resources may dictate, for example, that a state establish a tiered or staged approach in its monitoring design. This may involve an initial round of monitoring for a baseline set of indicators. A subsequent round(s) of targeted monitoring would follow for additional pollutants of concern.

10.1 What Indicators of Water Quality (e.g., physical, chemical, biological) Does the State Use as Baseline or Core Measures Statewide?

The objective in developing a baseline or core set of indicators for measuring attainment with WQS is not to limit monitoring programs to core indicators, but rather to identify a sound baseline for water quality assessment decisions. The core set of indicators includes physical, chemical, and biological measures of a waterbody. These indicators are appropriate measures of the ability of a waterbody to support its intended uses regardless of the degree of disturbance in the surrounding land use and watershed. Core indicators provide a scientifically valid foundation for consistent, practical, and cost-effective water quality assessments at the statewide level.

The Intergovernmental Task Force on Monitoring Water Quality (ITFM) identified potential indicators for describing water quality and presented rationales for their use in meeting water quality management objectives (ITFM 1995). Indicators included biological response and exposure, chemical response and exposure, physical habitat, and watershed-level stressors. The ITFM provided a general ranking of the indicators as high, medium, or low to describe the extent to which water resources support the uses designated under state WQS. The ITFM also stated that the appropriateness of an indicator for any given monitoring program would depend on the selection criteria, waterbody type, and management objectives.

Using the ITFM recommendations for water quality indicators as a starting point, Table 10-1 presents baseline or core indicators and supplemental indicators for water quality monitoring.

Core indicators are considered most important for measuring water quality for designated uses. Designated uses include aquatic life, recreation, public water supply, and fish and shellfish consumption. Core indicators could be used for initial water quality assessments and would be applied at both statewide and watershed scales. The core indicators should be supplemented with additional indicators based on the characteristics of the watershed, designated uses, and potential stressors (point and nonpoint sources) influencing the waterbody. Supplemental indicators might be used for followup monitoring to target the causes of water quality impairment or be included in the initial monitoring effort at a Statewide, watershed, or waterbody scale.

10.2 How Does the State Select Supplemental Indicators?

In addition to the core indicators listed in Table 11-1, supplemental indicators may be appropriate and should be included in the monitoring design framework as needed. This is particularly important for listing impaired waters needing total maximum daily loads (TMDLs) under section 303(d) of the Clean Water Act. Before a TMDL can be calculated, the pollutant or pollutants causing the impairment must be identified.

When selecting supplemental indicators, states should consider conditions that may cause or contribute to nonattainment of applicable WQS. For example, are there sources in the watershed that separately or collectively might contribute pollutants in amounts or combinations that could cause an exceedance of a water quality criterion, create toxic conditions, or accumulate in sediment or fish tissue? The following discussion presents basic considerations that may guide the process for determining the need for supplemental indicators for a monitoring design framework. Principal considerations include current and historical point sources, nonpoint sources, geology/hydrology, and land-use patterns. Other factors may include suspected pervasive pollutants such as those transported by atmospheric processes, or emerging pollutant concerns that the state might want to screen.

Table 10-1. Water quality indicators for general designated categories

Core and Supplemental Indicators				
	Aquatic Life & Wildlife	**Recreation**	**Drinking Water**	**Fish Consumption**
Baseline or Core Indicators **(Applied Statewide)**	*Condition of Biological Communities (EPA recommends the use of at least two assemblages) *Dissolved Oxygen *Temperature *Conductivity *pH *Habitat Assessment *Flow *Landscape conditions (*e.g.*, % cover of land uses) Additional indicators for lakes: *Eutrophic condition Additional indicators for wetlands: *Wetland hydrogeomorphic settings and functions	*Pathogen Indicators (*E.Coli*, enterococci) *Nuisance Plant Growth *Flow *Chlorophyll *Landscape conditions (*e.g.*, % cover of land uses) Additional indicators for lakes: *Secchi Additional indicators for wetlands: *Wetland hydrogeomorphic settings and functions	*Trace metals *Pathogens *Nitrates *Salinity *Sediments/TDS *Flow *Landscape conditions (*e.g.*, % cover of land uses)	*Pathogens *Mercury *Chlordane *DDT *PCBs *Landscape conditions (*e.g.*, % cover of land uses)
Potential Supplemental Indicators **(Applied at the watershed scale)**	*Toxicity *Nutrients *Hazardous chemicals in water column or sediment *Health of organisms	*Nutrients *Hazardous chemicals *Aesthetics	*VOCs (in reservoirs) *Hydrophyllic pesticides *Nutrients *Hazardous chemicals *Algae	*Other hazardous bioaccumulative chemicals

10.2.1 Point Sources in the Watershed

Point sources in the watershed may contribute pollutants that cause or contribute to nonattainment of WQS. Information about the type of facility and nature of discharges (e.g., process water or stormwater) contributes to an understanding of potential pollutants. Information about discharge characteristics should be available through permit applications and discharge monitoring reports. Many permittees are required to submit the results of a complete priority pollutant scan with their initial permit application and subsequent renewals. The permittee's file should also include compliance history information and wasteload allocation data and analyses. It is important to consider the potential cumulative impacts to a waterbody resulting from multiple sources of pollutants. Unless a TMDL has been completed for the waterbody, it is common for individual permits to be issued without consideration of other sources of regulated pollutants.

Point sources may have existed historically but may no longer be active. Historical sources may have contributed pollutants or contaminants to the environment that are still tied up within bed sediments in the waterbody or in soils at the site.

10.2.2 Nonpoint Sources in the Watershed

Nonpoint sources generally are related to land-use practices. Land use (e.g., rural, agricultural, urban, industrial) often dictates what indicators may be most suitable for water quality monitoring. Current and historic land-use practices in the watershed should be identified. Information about agricultural and animal husbandry practices, pesticide usage, urban/impervious surfaces, land management practices (e.g., forestry, mining), and best management practice (BMP) that would mitigate pollutant impacts should be reviewed. Past land-use practices may be very different from current practices, and residual pollutants may be present in the bed sediments in the water or in soils at the site. A discussion of pollutants associated with different land-use types and sources is presented in the third edition of Guidance for Assessing Chemical Contaminant Data for Use In Fish Advisories (U.S. EPA 1998) (see http://www.epa.gov/ost/fish) and in the ITFM Technical Appendix L, Ground Water Quality Monitoring Framework (ITFM 1997). Table 4-3 of the Guidance (U.S. EPA 1998) lists chemical contaminants by watershed type that bioaccumulate in fish tissue.

10.2.3 Geology and Hydrology

Geologic and hydrologic processes within and upstream from a watershed generally establish water quality conditions within the watershed. In some cases, weathering and transport processes for certain geologic areas may result in increased concentrations of metals, particularly arsenic, cadmium, mercury, and selenium. Increased concentrations may be found both in the water column and in underlying sediments (U.S. EPA 1998). Disturbances from land-use practices may aggravate already marginal natural water quality conditions.

10.3 How Do Core and Supplemental Indicators Fit Into the Monitoring Design?

Use of core and supplemental indicators may be integrated into a monitoring design framework in several ways. To illustrate the use of these indicators within a monitoring design framework, two simple frameworks are presented: staged monitoring and integrated monitoring. Chapter 11 provides a more detailed discussion of monitoring design frameworks and their advantages and disadvantages.

10.3.1 Staged Implementation of Core and Supplemental Indicators

An initial round of monitoring is conducted. Samples are collected for core indicators as appropriate for the WQS assigned to the waters from which the samples come. This round of monitoring is intended to assess the attainment/impairment status of waters represented by the sampling design. If a broad-scale, probability-based design is used for the assessment, then data collected during the initial round of monitoring are representative of all waters within the population from which samples were selected. These data provide a representation of the properties of waters that attain WQS or criteria as well as of those waters that are impaired. If a finer scale, targeted design is used, then these data represent the properties of the specific waterbodies or segments of waterbodies sampled and tested.

A second round of monitoring focuses on waters identified as impaired or having the potential to be impaired (based on analysis of ancillary data collected to help identify attributes of impaired waters). This round is focused on specific waters or waterbodies, so supplemental indicators are selected based on consideration of watershed characteristics and applicable WQS. During the second round of monitoring, these supplemental indicators are monitored at the sampling sites in addition to the core indicators. Further rounds of monitoring using supplemental and core indicators may be conducted, as appropriate, to better identify/delimit impaired waters, specific problems, and potential stressors or sources.

10.3.2 Integrated Implementation of Core and Supplemental Indicators

When the sampling framework is developed (statewide, watershed-specific, or waterbody segment-limited), appropriate core and supplemental indicators should be identified and included in the monitoring design. A monitoring design should include the core indicators appropriate for the designated uses and supplemental indicators based on consideration of watershed characteristics influencing each sampling location. Monitoring is conducted of all sampling stations specified in the monitoring design. Data from the monitoring may be used to assess the attainment/impairment status of waters represented by the sampling design. In some cases, analysis of samples for supplemental indicators may be delayed pending the results of first-tier indicator analysis. This may save money by reducing analytical costs associated with samples collected from waters where first-tier indicators show that water is attaining WQS.

10.4 References

Intergovernmental Task Force on Monitoring Water Quality (ITFM). 1995. The strategy for improving water-quality monitoring in the United States Technical appendix D: Indicators for meeting management objectives–summary and rationale matrices.

ITFM. 1995. The strategy for improving water-quality monitoring in the United States Technical appendix L: Ground water quality monitoring framework.

U.S. Environmental Protection Agency (U.S. EPA). 1991. Monitoring guidelines to evaluate effects of forestry activities on streams in the Pacific Northwest and Alaska EPA Region 10. EPA/910/9-91-001.

U.S. EPA. 1998. Guidance for assessing chemical contaminant data for use in fish advisories (see http://www.epa.gov/ost/fish).

11. Monitoring Network Design and Implementation Scenarios

Contents

11. Monitoring Network Design and Implementation Scenarios

This chapter emphasizes flexibility in designing and implementing approaches to monitoring that are based on well-documented science and judgment and that meet state objectives for both large-scale (e.g., statewide) and small-scale (e.g., drainage basin, waterbody) water quality characterizations. A state is likely to employ a combination of designs to meet its monitoring objectives. A state is also likely to use a variety of implementation strategies to balance the diverse demands of issuing NPDES permits, calculating total maximum daily loads (TMDLs), and assessing all state waters. The primary sections in this chapter follow:

- How does the state define its waters?
- What monitoring designs does the state use to support its decision making needs?
- How does the state implement its monitoring program to achieve its monitoring objectives?

11.1 How Does the State Define the Population of Waters Covered By Its Monitoring Design?

Defining the target population and developing the sampling frame are two key steps in designing a monitoring network. The target population is the set of waters that will be characterized by the monitoring design. In the broadest sense, the target population is all waters of the state; however, most states divide the population into subpopulations such as rivers, lakes, or wetlands. The target population is specified sufficiently well so that users of the monitoring information know what resources are included. The sample frame is a list or map that represents the target population or subpopulation.

The sample frame may be defined in various formats. For example, it may be a list of lakes within the state or a list of wellhead protection areas. It may be a map or GIS coverage of the river and stream network, wetland areas, groundwater aquifers, or coastal areas. When the sample frame is based on a map or GIS coverage, the scale of the coverage (e.g., 1:24,000 or 1:100,000) influences how well it coincides with the target population. Frequently, the sample frame does not coincide with the target population. It may include elements that are not part of the target population (overcoverage) or exclude elements that are part of the target population (undercoverage). The sample frame quality should be verified. This typically occurs during field monitoring when staff encounter a sampling location that does not meet the definition of the target population or subpopulation. Verification can also be conducted prior to identification of and visits to the sampling locations. The monitoring design, particularly probability survey designs, should address differences between the sample frame and target population.

The sample frame is used to generate the set of sampling units or sampling locations that is representative of the target population. The sampling units are the individual members of the target population that will be monitored to collect data about core and supplemental indicators. The water quality results from the sampling units are then used to describe the target population.

In water quality monitoring, as in most environmental monitoring, the target population has both spatial and temporal characteristics. Spatial characteristics are the fundamental consideration in monitoring network design, although temporal characteristics can be addressed when defining the target population. The temporal aspect can also be addressed through the selection of indicators and the definition of sampling protocols for each indicator.

11.2 What Are Monitoring Design Scenarios?

Monitoring design scenarios are part of the overall monitoring strategy that describes all aspects of monitoring activities, including defining the target population, selecting sample locations, selecting indicators to measure, defining how samples are collected and analyzed, and analyzing and presenting data and results. The state's assessment methodology should include a brief description of the monitoring designs used to generate water quality results. The methodology may refer readers to the state's monitoring strategy for additional details about the designs.

The following section focuses on options for designing the monitoring network. Three basic design options are described and may be used, typically in combination, to determine the extent of attainment or nonattainment of state waters and to describe the status of specific waters at an appropriate scale for supporting management decisions.

- Census—measurements are taken for every element in the population

- Judgmental (targeted)—measurements are taken at waterbody locations selected on the basis of experience of experts, and results are extrapolated on the basis of expert judgment

- Statistical survey (probability-based)—measurements are taken of a randomly selected sample of the target population, and statistical inferences about the distribution of values for the entire population are drawn from the sampled units.

States generally employ a combination of monitoring designs to address management questions. Judgmental designs tailored to a specific management question dominate state programs. Many states, however, are adding probability-based surveys. The U.S. Environmental Protection Agency (EPA) Environmental Mapping and Assessment Program (EMAP) program has assisted more than 30 states in designing probability surveys to meet a portion of their monitoring requirements.

11.2.1 When Does the State Census All Target Populations?

A census monitors each element of a population. This type of census design eliminates the error associated with not monitoring every member of the population. Census designs offer site-specific information about whether a given waterbody is attaining water quality standards (WQS) for all members of the population.

Because the total population of waters of a state is large, a census generally is impractical for comprehensive assessment. A census for all State waters would be expensive, time consuming, and resource intensive. Such censuses may be more appropriate for smaller populations or subpopulations such as designated bathing beaches or a specific drainage basin or waterbody segment. Even in these situations, a combination of designs is useful. For example, if all bathing beaches are monitored (census of bathing beaches), samples may be collected at random or judgmental locations from any given beach to characterize water quality at that beach.

Examples of a census used in water quality monitoring

No state currently employs a census for assessing all its waters, such as all lakes/ponds, all river/stream reaches, or all estuarine areas. Some states, however, apply a census to a subset of waters that share a particular designated use or feature (e.g., designated bathing beaches, drinking water reservoirs). The states may monitor all designated bathing beaches periodically during the swimming season, or all drinking water reservoirs.

11.2.2 When Does the State Use Judgmental (targeted) Designs?

Targeted designs rely on expert knowledge or best professional judgment based on experience to select sampling sites to meet specific objectives. These designs also use professional judgment to determine the spatial and temporal representation of each sample. By specifying the spatial extent of each sample location, analysts can piece together the total amount of waters assessed under a targeted design. Unless complete coverage can be pieced together, EPA recommends that judgmental designs cannot be used to draw conclusions about the extent to which the entire target population is attaining WQS. When judgmental designs are used, no quantitative statements can be made about the level of confidence in the sampling results (U.S. EPA 2000a).

Under a judgmental design, monitoring sites are selected in a nonrandomized way. Targeted site selection is generally used to answer specific questions regarding the condition of that particular site (U.S. EPA 1997a). Each site is selected on the basis of specific requirements that meet project or program objectives. The requirements may be environmental features (e.g., flow, wadeable riffled area), human population densities, and/or easy accessibility. A judgmental or targeted network is composed of sampling sites that are selected with a variety of criteria such as:

- Downstream discharge of a waterbody or watershed unit
- Known or suspected water quality problems
- Sources of potential water quality problems
- Upcoming events in the watershed, such as development or deforestation, that may adversely affect water quality.

Sites may also be selected to monitor best management practice (BMP) effectiveness or habitat restoration that is intended to improve waterbody quality. A judgmental network provides assessments of individual sites or reaches the sites are determined to represent (U.S. EPA 1997b). The extent or size of a waterbody that is represented by a given monitoring station will

depend on the waterbody type. Given the complexity of lakes and estuaries, no specific guidelines are available. For streams, a monitoring station can be considered representative of a stream waterbody for a distance upstream and downstream that has no significant influences that might tend to change water quality or habitat quality (U.S. EPA 1997a). More specific information on monitoring designs for wetlands has recently been released via www.epa.gov/owow/wetlands/monitor/#meth and is cited in U.S. EPA (2002).

Initially, the number of sites selected usually depends on the resources available. Often sites are prioritized, so high-priority sites are sure to be monitored if staff and funding resources are limited. In the long term, many sites may be added to the network over a number of years, so that the state is peppered with a dense sampling network. Each site may still be selected based on "core criteria." It is possible, however, that the network may be composed of sites that meet a variety of different criteria.

Judgmental sampling may be used in conjunction with other sampling designs to produce better documented results. In fact, judgment is an integral part of any monitoring network design. Existing information almost always helps develop a more efficient probability-based design.

Examples of judgmental designs

The water quality monitoring programs of most, if not all, states are based on a judgmental monitoring design. Several states have extensive fixed-station networks for chemical monitoring, macroinvertebrate sampling, and analysis of certain stream orders that were developed from targeted site selection. Depending on the sampling methods, some states monitor riffled areas only, whereas other states have both riffle and run sites.

The U.S. Geological Survey (USGS) gaging station network that is cost-shared with many states is a judgmental network. One of the best documented examples of a targeted design is the USGS National Water Quality Assessment (NAWQA) program. More detailed information on network design is documented in Design of the National Water Quality Assessment Program: Occurrence and Distribution of Water Quality Conditions.

Arkansas' basic monitoring program for rivers/streams consists of approximately 140 fixed monitoring sites that have been selected to assess the effects of point-source dischargers, assess the impact of nonpoint sources, monitor major rivers, and provide long-term chemical data for high-quality (least-impaired) streams by physiographic region for use in future WQS revisions. Sample sites have been selected to best represent monitoring objectives. An additional 40 monitoring stations are sampled yearly to address specific water quality interests, including unassessed major waters of the state. Through analysis of data collected at these sample locations, Arkansas was able to report on the condition of 10% of its rivers and streams in 1998.

Pennsylvania is implementing an extensive monitoring program for unassessed waters. Its design resembles a hybrid of judgmental and systematic random designs. State staff start at the mouth of a wadeable stream and move upstream, monitoring water quality at regular intervals approximately 8 to 10 miles apart. The monitoring location is selected based on the judgment of

the field crew. Factors influencing this selection include waterbody and land-use characteristics. This monitoring program focuses on aquatic life and uses a macroinvertebrate rapid bioassessment approach to indicate aquatic life use support. Other monitoring designs are being developed to monitor unasssessed nonwadeable rivers and lakes.

South Carolina has a fixed network of 314 integrator sites. These sites target the most downstream access of each of the Natural Resource Conservation Service (NRCS) 11-digit watershed units in the state, as well as the major waterbody types that occur within these units. For example, if a watershed unit ends in estuarine areas at the coast, integrator sites are located in both the free-flowing freshwater portion and the saltwater area. The results are used to establish trends and identify waters that may be quality limited. Intensive monitoring or other actions may be taken on the basis of the data from the integrator network. It is important to note that South Carolina also uses a probability-based design to provide a statewide overview of the extent that its waters attain WQS.

11.2.3 When Does the State Use Statistically Based (probabilistic) Designs?

States are increasingly incorporating statistical or probability-based monitoring into ambient water quality monitoring. Some of the numerous design options are summarized in this section. Details of each design, including conditions under which the design is appropriate, benefits and limitations of the design, and information on how to use the design, are included in USEPA QA/G-5S, Guidance for Choosing a Sampling Design for Environmental Data Collection (U.S. EPA 2000a). The document guidance, although general, does address many survey design issues that arise in aquatic monitoring. States, territories, and authorized tribes may also contact the EPA monitoring design team at http://www.epa.gov/wed/pages/EMAPDesign/index.htm to request technical assistance on designing statistical water quality monitoring networks.

The designs vary in complexity and offer a number of advantages for characterizing water resources with more precision and at different spatial and temporal scales. The simplest designs require the fewest sample locations to provide broad-scale descriptions of the extent that a population or subpopulation of waters is attaining WQS. These broad-scale characterizations satisfy the section 305(b) monitoring objectives to assess all waters and can provide an indication of the amount of waters that may be expected to be section 303(d) listed waters. More complex designs, requiring more sample locations, may also satisfy both the broader scale objectives and smaller scale monitoring objectives appropriate for developing management actions, including TMDLs.

All probability survey designs have these features:

- Reduce bias in the sample results by ensuring that sample units represent the target population

- Provide statistically unbiased estimates of the population mean, population proportions that pass or fail a standard, and other population characteristics

- Allow documentation of the confidence and precision of the population estimates.

Simple random sampling designs

Simple random sampling is the most basic probability-based design. It involves defining the target population and then using a technique to randomly select sample sites. For example, a simple random sample of a population of waterbodies can be taken by numbering all the waterbodies and then randomly selecting numbers from that list. Under this design, each element in the target population has an equal probability of being selected as a sample site. This method is the easiest way to generate an unbiased measure of the target population. The advantages of simple random sampling are that (1) sample size calculations are straightforward, (2) subsequent data analyses can use common statistical algorithms, and (3) the designs are easy to understand.

A potential limitation of simple random sampling is that it does not incorporate other auxiliary information. This may cause the monitoring design to be inefficient, in the sense that designs that do incorporate this information can have better precision. Auxiliary information can be nothing more than geographic location. Typically, survey designs that use geographic information are more efficient than those that do not.

Systematic and grid sampling designs with random start

In systematic and grid sampling, sample sites are selected at regularly spaced intervals over space or time. An initial location or time is chosen at random, and the remaining sampling locations are defined according to a regular pattern so that all locations are in defined, regular intervals along a linear feature, across an area (grid), throughout a volume, or over time (systematic). Systematic and grid sampling designs typically are used to search for hot spots and may be concentrated in a general location. They can be used to infer means, percentiles, or other parameters and also are useful for estimating spatial patterns or trends over time. These designs provide a practical and easy method for designating sample locations and ensuring uniform coverage of a site, population, or process. A limitation of most systematic or grid sampling designs is that exact precision estimates are not available without additional assumptions. An advantage is that all potential systematic samples are balanced across space or time.

Random tessellation stratified designs

Random tessellation stratified (RTS) designs incorporate features of simple random sampling and systematic sampling designs (Stevens 1997, Stevens and Olsen 1999). RTS designs use a two-step randomization process, with the first step being the random placement of a systematic grid over the region of interest and the second step being the random selection of units associated with grid cells. The RTS design guarantees a spatially balanced sample and enables precision to be estimated. If any spatial pattern in an indicator is present, the precision of the RTS design will be better than that of simple random sampling.

Designs incorporating additional information

Simple random sampling is most useful when the population of interest is relatively homogeneous. When the target population is not homogeneous, other techniques may be used to improve the precision of the sample. One of these techniques is stratified random sampling; another is called unequal weighting. Each improves the efficiency of the sampling design by incorporating additional information about the population. To achieve improved efficiency requires that the information be related to the indicator responses. Increasing the complexity of the sampling design also increases the complexity of the subsequent statistical analysis.

Stratified random sample designs. For stratified random sample designs, the target population first is separated into homogeneous subpopulations (strata). Sampling sites then are selected randomly from each stratum or subpopulation. The results of monitoring at each sample location within each stratum allow one to characterize both the target population as a whole and each stratum or subpopulation.

This design is particularly useful when the target population is heterogeneous (e.g., waters of the state) and can be organized into more homogeneous groups (e.g., low-order streams, high-order rivers, lakes, wetlands). By organizing the population into homogeneous strata, this design reduces the variability among the sample sites within each stratum and improves the precision of the results. Examples of different ways to stratify waters include:

- Waters stratified by waterbody type (e.g., streams/rivers, lakes, wetlands, estuaries)
- Waters stratified by other hydrographic discriminators (e.g., stream order, wadeable/nonwadeable, perennial/intermittent)
- Waters stratified by designated use (e.g., aquatic life, recreation)
- Waters stratified by ecoregion, physiographic province, elevation, land cover, or other geographic discriminator.

Within each stratum, simple random sampling, systematic random sampling, or RTS sampling may be used to select the sample. Stratified random sampling is also used to simplify operations or allow different sampling designs among the strata. For example, stratifying waters by waterbody type allows different designs to be used for streams, lakes, and wetlands. If a state uses a rotating basin approach to sample its streams, then the field work is operationally simplified by having the basins sampled in one year independent of those sampled in other years.

Unequal probability sampling. Unequal probability sampling can be viewed as a generalization of stratified random sampling and as a method for selecting a sample within a stratum. In unequal probability sampling, each unit in the population can have a different probability of being selected. In simple random sampling, each unit has the same probability of being selected. To implement unequal probability sampling requires auxiliary information for each unit in the population (technically each unit in the sampling frame). For example, if the target population is

streams, it is likely to include many more first-Strahler-order streams than fourth-Strahler-order streams. Unequal selection may be used to ensure that the randomly selected sample results in approximately the same number of sample units in each Strahler order. By not treating the subpopulations as formal strata, the unequal weighting technique allows researchers to characterize subpopulations that are not based on Strahler order. Unequal probability sampling can increase the precision if the auxiliary information is strongly correlated to an indicator.

Additional design options. The sample designs discussed represent those that are commonly used. Many other types are available, including cluster sampling, multistage sampling, multiphase sampling, ranked set sampling, adaptive cluster sampling, and survey designs over time, to name a few. Which design to use depends on the monitoring objectives, the type of aquatic resource to be monitored, and the information available about the population (specifically, available sampling frames). As an example, it is possible for a sampling design to provide a statewide sample for perennial streams and rivers and a more intensive sample of perennial streams and rivers in a specific subregion. Another example would be a sample for the entire Puget Sound in Washington with intensified sampling in harbors or bays of specific interest.

Examples of statistical (probability-based) sampling designs

South Carolina uses a statewide probability design network in addition to the fixed station integrator network described earlier. The probability network provides data to make inferences, with known statistical confidence, about the condition of the state's water resources. The design is stratified by waterbody type: streams, lakes/reservoirs, and estuarine resources. Within each stratum the state uses unequal weighting to ensure sufficient sampling across subpopulations. Each sample location is monitored annually for indicators of biological condition, habitat, and sediment quality. The state conducts more frequent monitoring for specific physical/chemical indicators.

Maryland's biological stream survey monitoring program is designed to provide a statistically unbiased estimate of the condition of wadeable (first- through third-order) nontidal streams and rivers. Basins in the state are divided into three geographic regions for assessment purposes. From 1995 to 1997, basins in one geographic region were sampled each year, and one basin in each region was sampled twice. Random monitoring site selections were made from all sections of streams that could physically be sampled. Approximately equal numbers of first-, second-, and third-order streams were sampled. The number of sampling sites of each stream order in a basin was proportional to the number of miles of the stream order in the state. The sampling strategy was designed to allow Maryland to develop statistically valid estimates of largemouth bass densities, miles of streams with degraded physical habitat, and miles of streams with poor Index of Biological Integrity scores. Other types of monitoring programs (generally judgmental) apply to lakes, larger order streams, and wetlands and estuaries.

Historically, West Virginia has employed a targeted water quality monitoring program for streams/rivers to focus on known or suspected pollution problems. The state is divided into 32 watersheds (8-digit HUC), which are organized into 5 watershed groups. One watershed group

is monitored in each year of the state's 5-year rotating assessment monitoring cycle. A list of streams for the watershed group is developed from EPA's Water Body System database, and samples are selected from as many listed streams as possible close to the mouth of the stream. In 1997, West Virginia established a random monitoring program to complement the targeted stream program. Approximately 30 to 45 stream locations within a watershed group are selected randomly from an EPA database to allow development of statistical comparisons among watersheds.

Nebraska employs a rotating basin approach for water quality assessment monitoring. The monitoring strategy targets resources in two or three river basins annually to allow for intensive efforts to increase the identification and abatement of pollution problems. Monitoring is conducted for rivers/streams, lakes, and wetlands located within the selected basins. All 13 water basins in Nebraska are monitored over 5 years. Since 1997, Nebraska's biological monitoring program has been based on a probabilistic methodology developed in association with EPA. Approximately 40 biological monitoring sites are selected randomly each year from the perennial streams within the water basin of interest for that year. Sample sites for other monitoring purposes are selected to best represent monitoring objectives and are based on professional judgment. The approach also supports coordination and integration of environmental programs through a Basin Management Approach.

Indiana employs a 5-year rotating basin (eight-digit HUCs) approach to monitoring surface waters. The strategy includes fixed stations, randomly selected stations for biological and water chemistry monitoring, pesticide monitoring, bacteriological monitoring, NPDES permit monitoring, TMDL development monitoring, and targeted fish and surficial aquatic sediment monitoring. The monitoring strategy is designed to describe the overall environmental quality of each basin and to identify impaired waters.

11.3 How Does the State Integrate Broad-Scale Monitoring Designs With the Need To Identify Site-Specific Impairments?

Some designs address this issue specifically; others require a supplemental or nested approach to achieve monitoring objectives at a broad-scale statewide level and at a small-scale site-specific level. The census design provides data for both broad-scale (e.g., statewide, waterbody type, designated use) and site-specific (e.g., waterbody segment) characterizations. A judgmental design may provide characterizations at both levels if complete representation of all waters is achieved, but judgmental designs have historically provided site-specific information only for a limited amount of waters. The probability designs most frequently used for water quality monitoring are designed to provide statewide or watershed-level characterizations, though they do provide site-specific data for individual sample locations.

Probability designs support site-specific descriptions of water quality in a variety of ways. Some states use the probability design to systematically sample new locations each year. Under this approach the state gets site-specific information on waters that had previously been unmonitored, in addition to broad-scale characterizations of water quality throughout the state or watershed.

Another way probability designs support identification of impaired waters is by providing information to refine future probability designs. When analyzing the results of the probability survey, staff may identify associations between impaired waters and certain land-use, source-type, or other characteristics. Future probability (or judgmental) designs can intensify sample collection around these land uses or other characteristics in an effort to identify, at a smaller scale, waters that are impaired.

Intensified monitoring to characterize smaller scale areas can also be incorporated into the probability design from the beginning. A state may have existing data or information that certain areas are likely to be impacted (or likely to be pristine). To confirm that information, the broad-scale probability design can be supplemented with intensified sampling to characterize smaller scale conditions.

If followup monitoring is required after a probability survey to obtain site-specific information, either a judgmental or probability approach may be used. In either case, professional judgment will improve the precision and reduce the cost of the followup monitoring. The type of information used to determine where to monitor and what indicators to monitor includes:

- Results from probability and judgmental monitoring
- Predictive models
- Preexisting data and information on water quality conditions
- Land-use data
- Analysis of potential sources in the watershed (see discussion on selecting supplemental indicators in Chapter 11).

11.4 How Does the State Implement Monitoring Designs To Achieve Its Monitoring Objectives?

States generally structure implementation to distribute monitoring resources among multiple monitoring objectives and associated management priorities. There will be some shifting from year to year as priorities shift among TMDL calculation, WQS revision, NPDES permit issuance, and so on. However, to ensure that resource allocations are appropriate, it is critical that states maintain sufficient funding for a base monitoring network. This network provides the foundation for other monitoring priorities by describing the extent to which waters are attaining WQS and characterizing the causes and sources of impairments. A probability design is the most effective way to maintain a base network and maximize the resources available to support other, more narrowly targeted monitoring priorities.

States, territories, and authorized tribes should describe, in a monitoring strategy and periodic monitoring plan, both the monitoring designs and implementation plans to achieve monitoring objectives. This section explores implementation options for network designs that achieve the monitoring objectives of a state, territory, or tribal water quality management program Network designs may be tailored for waterbody types, individual basins, and applicable designated uses.

11.4.1 Tailoring Monitoring Designs for Waterbody Types

States commonly organize monitoring designs according to the waterbody types present in the state. The specific design could be a basic stratified random design, a stratified design combined with unequal weighting, a stratified design with intensified sampling in areas of concern, a ranked set, or judgmental. Regardless of the design selected, tailoring it to waterbody types requires expert professional judgment to organize different waterbody types into relatively homogenous groups or strata. The text box shows an example of potential strata for surface waters.

Example Strata for Surface Waters

Rivers/Streams
- Intermittent streams
- Wadeable streams
- Nonwadeable streams/deep rivers

Lakes
- Small lakes (<50 acres)
- Medium-sized lakes (50-250 acres)
- Large lakes (>250 acres)

Wetlands
- Depressional wetlands
- Slope wetlands
- Fringe wetlands
- Flats

Estuaries
- Small estuaries (<250 km^2)
- Large estuaries ($\geq$250 km^2)
- Large tidal rivers
- Small tidal rivers

Depending on the resources available to the effort and the other monitoring objectives that need to be accommodated, this design may be implemented under a variety of schedules. Ideally, the design would be implemented statewide for all waters every year. Alternatively, the jurisdiction could implement the design for one or more strata each year to achieve comprehensive assessment over several annual monitoring cycles (see text box).

11.4.2 Tailoring Monitoring Designs for Individual Basins

Rotating basin monitoring designs focus monitoring and programmatic activities and resources in a few watersheds or basins (e.g., eight-digit hydrologic unit codes) each year, because

Potential Schedule for Implementing Waterbody-Type Monitoring Design

Year 1 - Wadeable streams
Year 2 - Nonwadeable rivers
Year 3 - Lakes
Year 4 -Wetlands
Year 5 - Estuaries and coastal waters
Year 6 - Intermittent streams

available resources preclude comprehensive monitoring and programmatic activities across the state. Over time, typically a 5-year cycle, all basins are monitored. Then the cycle begins anew.

The state may use any combination of the designs described in Section 11.2, or other appropriate designs, to meet its monitoring objectives within each basin.

11.4.3 Tailoring Monitoring Designs for Applicable Designated Uses

States, territories, and authorized tribes need to implement monitoring designs that describe the extent to which all applicable uses of state waters are attained. This may be achieved through the selection of indicators that are measured at each sample location. It may be achieved by developing a specific monitoring design tailored to the designated use. Or it may be achieved by a combination of options. For example, a small-scale probability or judgmental design may be used to assess designated uses that apply to a relatively small subset of waters such as high-use beaches, shellfish areas, or public water supplies. Meanwhile, other designated uses may be assessed as part of a multipurpose monitoring design that includes an appropriate suite of indicators.

Aquatic life

Most monitoring designs focus on aquatic life use. This designated use generally applies to all waters of the state, unless a use attainability analysis has demonstrated that the aquatic life use was not attainable.

States increasingly are building a probability design into their monitoring program and using it to describe the extent to which aquatic life uses are supported throughout the state or watershed. Many states are in the early stages of this design and have only developed it for a few waterbody types, typically wadeable rivers and streams. Some states have expanded this design into lakes, coastal waters and wetlands. Over time, these designs should address all waterbody types.

The results of the probability survey serve as an objective check on the results of judgmental designs. As described in Section 11.3, the results of the probability design can also be used to locate impaired waters needing restoration or pristine waters needing protection.

Recreation

A combination of monitoring designs and scales that are tailored to the level of recreational use assigned to the waterbody may be most appropriate for assessing recreational use and supporting beach advisories and closures.

For those areas identified as bathing beaches, the sampling design should provide enough information to determine whether to issue a public health-based advisory or closure. Even within beach areas, states may employ different levels of sampling intensity. This would allow the allocation of monitoring resources to target the high-use and/or high-risk beach areas.

For an assessment of whether all recreational waters (i.e., those designated for primary contact recreation) are attaining their uses, states may want to rely on probability design to evaluate the extent to which all other waters are attaining WQS.

Fish consumption

For fish consumption use attainment, the state will generally identify the waters with high use for fishing and monitor fish species that are likely to pose health threats if consumed. The simplest sampling design may be simple random or stratified random.

Shellfish consumption

EPA recommends following the National Shellfish Sanitation Program monitoring protocols to assess attainment of the shellfish consumption use. States implementing this program typically use simple random sampling at all approved shellfish harvest areas. State water quality agencies should coordinate with the responsible state agencies to maximize use of data collected under the shellfish sanitation program and minimize additional monitoring.

Public water supply

In addition to the statistical designs in Section 11.2.3, ranked set or cluster designs may also be appropriate to monitor waters designated as public water supplies (U.S. EPA 2000a). This is particularly true for supplies for which the state has already developed source water assessments. Source water assessment reports, which provide detailed information about the potential causes and sources of contamination to public water supplies, serve as valuable information in designing the ranked set or adaptive cluster designs and selecting the most appropriate indicators to measure.

11.5 References

Stevens DL, Jr. 1997. Variable density grid-based sampling designs for continuous spatial populations. Environmetrics 8:167-195.

Stevens DL, Jr., Olsen AR. 1999. Spatially restricted surveys over time for aquatic resources. J Agric Biol Environ Stat 4:415-28.

U.S. Environmental Protection Agency (U.S. EPA). 1997a. Guidelines for preparation of the comprehensive state water quality assessments (305(b) reports) and electronic updates: Report contents and supplement. September 1997. EPA/841/B-97-002A & B.

U.S. EPA. 1997b. Monitoring guidance for determining the effectiveness of nonpoint source controls. September 1997. EPA/841/B-96-004.

U.S. EPA. 2000. Guidance for choosing a sampling design for environmental data collection (QA/G-5S). Draft document.

U.S. EPA. 2002. Methods for evaluating wetland condition: Study design for monitoring wetlands. Office of Water; Washington, DC. EPA 822-R-02-015.

APPENDIX C (DRAFT)

STATISTICAL CONSIDERATIONS FOR DATA QUALITY OBJECTIVES AND DATA QUALITY ASSESSMENTS IN WATER QUALITY ATTAINMENT STUDIES

Michael Riggs, Dept. Statistical Research, Research Triangle Institute

Acknowledgements:

We are grateful to Andy Clayton (RTI), Florence Fulk (EPA/NERL), Laura Gabanski (EPA/OWOW), Susan Holdsworth (EPA/OWOW), Forrest John (EPA/Region 6), and Roy Whitmore (RTI) for providing thoughtful reviews of earlier drafts of this appendix and for suggesting a number of changes which greatly improved the final version.

Appendix C Table of Contents

Appendix C Table of Contents (continued)

Appendix C
Statistical Considerations for Data Quality Objectives and Data Quality Assessments in Water Quality Attainment Studies

C.1 The Data Quality Objectives Process: Principles of Good Study Design

C.1.0 Introduction to DQO Procedure

The process of determining if a body of water meets water quality standards can be divided into two phases. The first phase, the study design phase, encompasses seven activities involved in specifying the appropriate research questions and developing a strategy for collecting the data needed to answer them. The seven activities comprise the Data Quality Objective (DQO) process. Part C.1 of this appendix details the four "statistical" DQOs (i.e., DQOs 4-7) and provides guidance for choosing the appropriate statistical tools for achieving them. The second phase, called the Data Quality Assessment (DQA), includes all of the statistical procedures necessary to answer the questions of interest in the face of the uncertainty inherent in the data and the data collection methods. Parts C.2 and C.3 of this appendix provide guidance for selecting and implementing inferential statistical techniques needed to complete a DQA to support a water quality attainment decision. Part C.2 introduces a collection of techniques called exploratory data analysis (EDA) that is useful for determining structure and pattern in the data and for examining the validity of various assumptions that may underlie the more formal processes of statistical inference (i.e., interval estimation and hypothesis testing) that are developed in Section C.3.

C.1.1 Review of the Basic DQO process

EPA defines environmental data to include data collected directly from field measurements or compiled from existing databases or literature. When such data are used to select between two alternative conditions (e.g., attainment or non-attainment), EPA requires that organizations responsible for monitoring ensure that the data used to characterize the environmental processes and conditions of concern are of the appropriate type and quality for their intended use and that environmental technologies are designed, constructed and operated according to defined expectations. The Data Quality Objective (DQO) process is EPA's recommended tool for developing sampling designs for data collection protocols that will ensure these conditions such that the quality of the data are sufficient to support decision-making within tolerable levels of decision error. The 7-step DQO process is described in detail in the EPA document, *Guidance for the Data Quality Objective Process* (EPA/600/R-96/055). The steps are:

1. Define the problem
2. Identify/state the decision which must be made
3. Identify the information (data) needed to make the decision
4. Define the target population
5. Develop a decision rule and the population parameters needed to make the decision
6. Specify the tolerable limits of error in making the decision
7. Choose an optimal sampling design

Step one involves the formulation of the general problem and a corresponding conceptual model of the hazard to be accessed. For example, we may want to examine a particular reservoir to estimate the mean selenite concentrations in the water column. Having identified the problem, it is then necessary to select a project manager, technical staff and the associated resources that will be needed to collect the appropriate data from the target waters.

In Step 2 of the DQO process, the water quality attainment (WQA) question and the alternative actions to be taken in response to the answer, must be clearly stated. The question will often be as simple as, "Does this reservoir attain the current WQA criterion for selenite concentration in the water column?" Alternative actions, may include listing the reservoir and/or instituting other remedial activities or use restrictions. The WQS question and the alternative action(s) together comprise the "decision statement" which is critical for defining decision performance criteria in Step 6 of the DQO process.

In Step 3, the specific field data that are needed to resolve the decision statement are identified, as well as any pertinent data in the literature or historical databases. At this time, the appropriate field and laboratory measurement and analytical methodologies should also be identified. The specifications and criteria for these measurements, measurement devices, laboratory procedures, etc. may be refined later in the DQO process.

Steps 1-3 require consultation among the planners and regulators and are primarily concerned with identifying qualitative criteria for the study. Once these have been established, the remaining 4 criteria must be addressed through consideration of statistical principles for population estimation.

C.1.2 Defining the Target Population

The target population is the entire set of elements to which the investigators desire to extrapolate the findings of their survey or monitoring design. Thus depending on the study objectives (DQOs 1 & 2), the target population may be all the waters in a state, a specific type or class of waters in a state, all the waters within a specific drainage, or only the waters in a particular stream reach or pond. Additional background on monitoring designs and target populations are presented in Chapter 12 of this document.

For environmental studies, it is critical that such populations are bounded in space and time. For example, investigators may desire to quantify the condition of a specific stream-segment in January of a specific year. Whereas in most biological studies the population elements, generally called sampling units, are discrete individuals (e.g., fish or people), in water quality studies (WQS), they are often volumes of water. Thus, if a **sampling unit** were defined as a 1-liter aliquot, the target population for a January, 2001 water quality assessment of a 3-mile reach of a stream would contain all of the 1-liter quanta, which flowed through the stream segment during January of 2001. Estimates from such a sample are only valid for inferences about that particular 3-mile reach in January 2001. They cannot provide a statistical basis for inferences to the same stream reach at any other time (e.g., August 1999), nor for other reaches of the same stream or for different streams.

C.1.3 Developing a Decision Rule and Choosing Parameters and their Sample Estimators

The population factor of interest is called a **parameter**; e.g., the proportion of sampling units in the target population which exceed some standard. The surest way to measure a population parameter is to make the measurement on every member of the population. The process of doing so is called a **census**. While this may be feasible for a small closed population (e.g., all the palm trees on a small island), it is not likely to be so for most bodies of water. Typically, some subset of the population will be selected as representative of the target population and the measurement will be made from this subset or **sample**. The measurement obtained from the sample (e.g., the sample proportion) is called a **statistic**; it provides the investigator with an estimate of the more difficult and expensive population parameter. To avoid confusion, in this appendix "sample" will be used in the statistical sense; i.e., it will refer to the collection of the individual volumes taken from the target water. The volumes themselves (e.g., 1-liter aliquot) will be referred to as "sampling units".

The choice of the criterion for whether a water body or stream segment has attained water quality standards or is impaired will generally be based on prior scientific investigations. Although such criteria exist for biological, physical, and chemical components of an aquatic system, this appendix focuses on the chemical components. For toxic (e.g., arsenic) pollutants, non-priority chemicals, and physical parameters (e.g., pH), EPA has established two types of criteria: acute and chronic. Acute criteria are based on 1-hour mean chemical concentrations determined from laboratory studies. The acute hourly means are called Criteria Maximum Concentrations (CMC) and have been tabulated for both toxic and non-priority pollutants (EPA 1999; http://www.epa.gov/OST/standards/wqcriteria.html). EPA regulations specify that the acute criteria for a pollutant must not be exceeded more than once in a given 3-year period for any water body. Chronic criteria are based on 4-day (i.e., 96-hour) chemical means, called Criterion Continuous Concentrations (CCC), obtained from laboratory and/or field studies. EPA regulations require that (1) the 30-day mean concentration of a pollutant in a body of water must not exceed the CCC for that pollutant more than once in a 3-year period and (2) no 4-day mean can exceed 2×CCC in any 4-week period. The limiting concentrations, durations and 3-year frequencies specified in the criteria are based on biological, ecological and toxicologic modeling studies and have been designed to protect aquatic organisms and ecosystems from unacceptable effects. Details regarding the calculation and scientific bases for CMC and CCC have been documented by Stephan et al. (1985).

Statistical methods must be employed to determine if the acute and/or chronic criteria for parameters such as dissolved oxygen, pH, surface temperature, etc. have been exceeded. For example, it may be known that concentrations of chemical X greater than 10 µg/l represent an important threshold for algal blooms. Given such a threshold, it is then necessary to define a sample statistic to which it may be compared. The choices may include: the maximum tolerable proportion of sampling units in the sample that exceed the threshold (e.g., 10%), the sample mean concentration, the sample median concentration or some percentile (e.g., the 95^{th}) of the sampling unit concentrations in the sample. The choice of which statistic is "best" depends on the expected behavior of the sample statistics and on the study objectives. For example if the distribution of the concentrations among the sampling units is expected to be bell-shaped (or approximately so) then the mean may be the best choice. However, if the distribution is

expected to be skewed (e.g., most sampling units have low concentrations but a few have very high concentrations), then the median might be preferred. On the other hand, the investigators may wish to make their decisions on the basis of extreme values, however rare in the sample, in which case the 95th percentile might be the appropriate statistic.

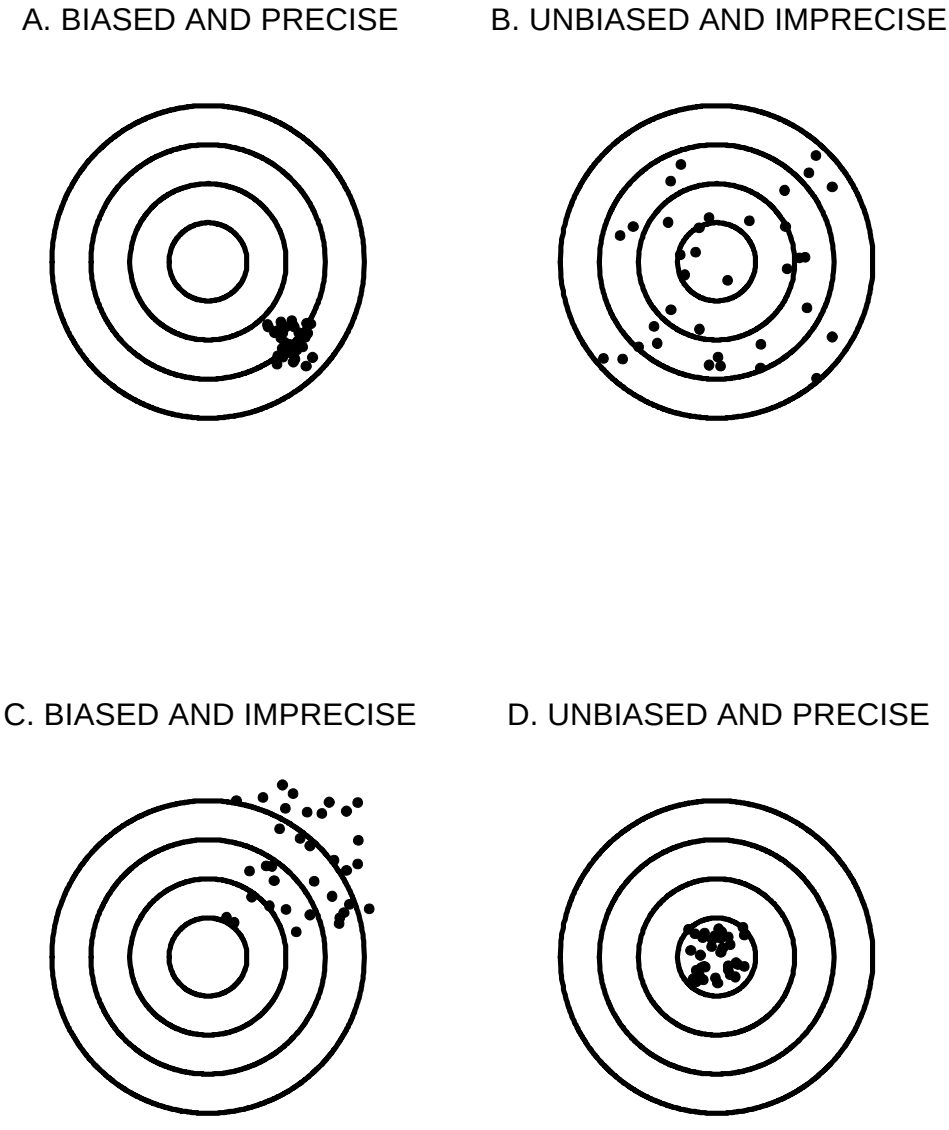
The choice of whether to measure central tendency or extreme values in a population distribution usually depends on whether we want some convenient means of characterizing the "average" condition in the population or if we want to know the magnitude of the "best" or "worst" conditions in the population. In the former situation we would focus on the mean or median. For example if some sort of remedial action had been applied to reduce nutrient loading in a body of water, we might want to compare mean or median concentrations of various phosphates or algal abundances before and after treatment to determine if there had been a "general" improvement in the body of water. Alternatively, if we had human health concerns associated with consumption of fish tissue containing mercury above some threshold concentration, we might want to estimate the 95th percentile of tissues concentrations of mercury in the resident fish population. The reason for this is that if 5% or more of the fish have tissue levels above a critical threshold for human health effects, there would be at least a 1/20 chance of toxic exposure due to human consumption of those fish. This would be so regardless of the magnitude of the population mean or median mercury concentration. In other words, interest focuses on the extremes (e.g., the 95th percentile) because it is only the extremes that are likely to effect human health.

C.1.4 Bias, Imprecision and Decision Error

Decisions in WQS will nearly always be based on sample data. Because sample data are subject to both bias and imprecision, decisions based on such data will be subject to error. To understand how to control this error, one first needs to understand the nature of bias and imprecision in sample-based inference.

It is intuitive that not every sample estimate is good, in the sense that its value is close to that of the unmeasured population parameter. In particular, the sample statistic may be biased, imprecise or both. Figure 1 illustrates four possibilities as distinct shot patterns around a bull's eye; each bull's eye represents a population target parameter, while the shots are statistical estimates of the parameter obtained from repeated sampling of the target population. Figure 1a illustrates the case where the statistical estimates are biased but precise. Sampling **bias** results from systematic error caused by sampling which favors some individuals over other population members. For example radio talk-show call-in surveys tend to over-represent disgruntled listeners. Thus their responses deviate from the true population parameter in the same direction (i.e., are biased) and tend to be more alike. This homogeneity is reflected by the tight clustering of the shots (i.e., the shot pattern is precise) in Fig. 1a. The statistical estimates in Fig. 1b are not skewed in any particular direction, thus there does not appear to be any bias. However the pattern is highly dispersed around the true parameter value. This is indicative of considerable heterogeneity in the target population, which leads to **imprecision** in the sample statistics. Imprecision and heterogeneity are reflected in increased dispersion of the statistical estimates

Fig. 1. Bias and precision as represented by shot patterns on a target. Each bulls-eye represents the true target population parameter and the shots represent sample estimates of the target parameter.



about the value of the target parameter. In situations like Fig. 1b, a large number of shots (i.e., samples) are needed to ascertain the approximate center (i.e., true parameter value). Whereas Fig. 1c depicts the worst case scenario (biased and imprecise), Fig. 1d illustrates the best case (unbiased and precise). Note that in Fig. 1d, a much smaller number of samples are needed to locate the bull's eye than in Fig. 1b; in fact, a parameter estimate based on any one of the samples in Fig. 1d will be correct, while one would need to average over all of the samples in 1b to get a parameter estimate that was close to the true target value.

Water quality attainment decisions are made on the basis of probabilities derived from the sample estimates by the application of statistical inferential procedures. Thus, bias or imprecision in the sample estimates may lead to erroneous probability statements. The computed probabilities support the acceptance or rejection of two competing hypotheses: the **null hypothesis** (H_0) that the water attains WQ standards vs. the **alternative hypothesis** (H_a) that the water is impaired. Given these two competing hypotheses, imprecision and/or bias in the sample estimates of the desired population parameters creates the potential for two types of decision errors.

So-called **Type I errors** occur when the null hypothesis is incorrectly rejected; i.e., a water that attains WQ standards is erroneously judged to be impaired. A **Type II error** occurs when an impaired water is erroneously judged to be in attainment. Type I and II errors are often compared to the judicial errors of convicting an innocent man (Type I) and letting a guilty man go free (Type II). Notice the near inevitability that when we decrease the probability of one error we coincidentally increase the probability of the other. Thus a subjective decision is usually made to guard against one at the expense of the other. For example, the U.S. judicial system has traditionally protected against convicting the innocent at the expense of letting the guilty go free. The Stalinist Soviet Union took the opposite view and jailed a large proportion of its population, innocent and guilty alike, with the result that nearly all criminals were sent to labor camps and Soviet crime rates were exceedingly low. Similar to the U.S. judicial system, the scientific community has focused on protecting against Type I errors rather than Type II errors. This is apparent in the published scientific literature wherein Type I error rates (the α-level) are almost always reported, while the Type II error rates (the β-level) are much less commonly considered.

The statement about the near inevitability of increasing β whenever α is reduced is conditional. To be strictly correct, we must say that *for a given quantity of evidence* (i.e., the sample size), decreasing α will inevitably increase β. This suggests that it is indeed possible to decrease α and β simultaneously, if one increases the amount of evidence. In the statistical assessment of water quality attainment issues, the sample is the ultimate source of the evidence. Thus, if we increase the number of sampling units (n) in the sample until it equals the population size (N), α and β will decline to zero. Unfortunately, sampling units are expensive to collect and process. Consequently the cost of simultaneous control of α and β to low levels (e.g., 0.05) is generally prohibitive, often requiring hundreds or even thousands of sampling units per sample.

C.1.5 Quantifying Sampling Error: Confidence Intervals

Clearly the Types I and II decision error rates reflect the imprecision in the sample estimates. The imprecision in the sample estimate, called **sampling error,** is due to approximating the true

population parameter (e.g., the mean) with an estimate computed from a sample. The **standard error** (SE) of a sample estimate of a population parameter provides a quantitative expression of the sampling error. For example, the standard error of the sample estimate $(\bar{x})$ of the population mean (μ) is computed as shown in Box 1. Notice that as the sample size (n) increases to values which approach the population size (N), the correction factor (fpc) goes to zero and, further, when fpc=zero, the standard error will also be zero. This provides a mathematical justification for our earlier statement that the decision error rates, α and β, will be zero whenever the population is censused. Another way of saying this is that sampling error only exits when a population is sampled. As previously stated, in the context of water quality attainment studies, the sample size (n) will always be extremely small relative to the population size (N). In such cases, the FPC is essentially equal to 1.0 and therefore is ignored. However this suggests that the standard errors of estimates obtained from water quality attainment sampling efforts are not trivial and hence the potential for substantial Type I and II decision errors must be addressed in the DQO process (step 5).

One way that statisticians deal with sampling error is to construct **confidence intervals** about the sample estimates such that the interval has some known probability (e.g., 95%) of containing the true population parameter. The confidence interval is a statement about the confidence we have in the sample estimate of a population parameter, θ. Algebraically, this statement is written as shown in the first expression in Box 2. By convention, the **confidence level** is expressed as a percent (e.g., 95%). For example if we desire to hold the Type I error rate to α=0.05, we are in effect saying that there is a 5% chance that our estimate is incorrect; thus, the corresponding confidence interval says that we can be 95% confident that the value of the unknown population parameter is within the bounds of the confidence interval. A more mathematically rigorous explanation of the 95% confidence interval is as follows: if one drew 100 different random samples from the population and computed 95% confidence interval estimates of the population parameter θ, from each sample, we would expect that 95 of the computed confidence intervals would include θ. Because this interpretation of the confidence interval is based on the concept of repeated sampling of the target population, this approach to statistical inference is called **frequentist statistical inference**.

Similar to hypothesis tests (see Section C.3.1), confidence intervals may be two-sided or one-sided (Box 2). The first expression in Box 2 is a two-sided 1-α confidence statement. Upper and lower 1-α one-sided confidence statements have the general form shown in the second and third expressions in Box 2. Two-sided confidence intervals are appropriate when one desires a sample estimate (with $100\times(1-\alpha)$% confidence) of an unknown population parameter, as is the case in most investigations of processes that characterize natural populations and/or ecological systems. One-sided confidence intervals are appropriate when one wants to compare the sample estimates to a specific regulatory standard. For example, a lower one-sided confidence interval would be appropriate for comparing an observed chemical concentration to the maximum allowable concentration of that chemical. Conversely, an upper

Box 1-a: The Sample Mean and Its Standard Error

Suppose X represents the population factor of interest. Let *N* denote the population size, μ the population mean, and s^2 the population variance. Let $X_1, X_2, \ldots, X_n$ represent *n* data points, i.e. a random sample of *n* unit from the population of *N* units.

The sample estimate of the population mean μ is the sample mean $\bar{x}$:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

The standard error $\bar{x}$ is given by:

$$SE(\bar{x}) = \sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)} = \frac{s}{\sqrt{n}} x \sqrt{1-\frac{n}{N}}$$

where s is called the population standard deviation and the quantity $1-\frac{n}{N}$ is the finite population correction factor (fpc). Oftentimes, the population variance s^2 is unknown but may be estimated using the sample variance s^2:

$$s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{x}^2\right)}{n-1} .$$

Hence, the standard error of $\bar{x}$ may be estimated by replacing s^2 with s^2, or with s, the sample standard deviation.

Box 1-b: Example for Calculating the Sample Mean and its Standard Error

Consider a random sample of 10 of the 244 monthly turbidity measurements taken from the Mermentaut River between June 1980 and April 2000. The measurements (in NTU) were: 34, 58, 87, 145, 14, 38, 62, 95, 160, 320.

$$\bar{x} = \frac{34+58+87+145+14+38+62+95+16-+320}{10} = \frac{1013}{10} = 101.3$$

$$s^2 = \frac{(34-101.3)^2+(58-101.3)^2+(87-101.3)^2+\ldots+(160-101.3)^2+(320-101.3)^2}{10-1}$$

$$= \frac{4529.29+1874.89+204.49+\ldots+3445.69+47829.69}{9} = \frac{73006.10}{9}$$

$$= 8111.7889$$

$$s = \sqrt{8111.7889} = 90.07$$

Note that an infinite number of possible turbidity measurements could have been taken from throughout the target area during the 20-year monitoring period. In these circumstances the finite population correction factor is essentially 1.0 and can be ignored. Thus the estimate of the standard error is $\bar{x}$ is simply:

$$\hat{SE}(\bar{x}) = \sqrt{\frac{8111.7889}{10}} = 28.48\,.$$

Box 2: General Forms for 100 x (1-a)% Confidence Intervals

Let ? denote a population parameter. A two-sided 100x (1-a)% confidence interval about a sample estimate of ? has the general form:

$$\Pr\left[a_1 \le q \le a_2\right] = 1 - a$$

where 1-a = the desired confidence level
a_1 = the 1-a/2 lower bound of the sample estimate of ?
a_2 = the 1-a/2 upper bound of the sample estimate of ?.

One-sided confidence intervals have the general form

$$\Pr\left[-k \le q \le b_2\right] = 1 - a \quad \text{for an upper one-sided confidence interval}$$

$$\Pr\left[b_1 \le q \le +k\right] = 1 - a \quad \text{for a lower one-sided confidence interval}$$

where -k = 0.0 for proportions and variances; -8 for means and medians
+k = 1.0 for proportions; +8 for means, medians and variances
b_1 = the computed 1-a lower confidence limit
b_2 = the computed 1-a upper confidence limit.

The upper and lower confidence limits a_1, a_2, b_1 and b_2 and functions of the desired confidence level and the sampling distribution of the sample statistic used to estimate ?.

one-sided confidence interval would be appropriate for comparing the observed abundance of an organism to the minimum abundance deemed necessary for survival of the species and/or for the health of the ecosystem. Because upper and lower 1-sided confidence intervals are easily derived from the corresponding 2-sided confidence intervals, only the latter will be presented in this appendix. In nearly all cases, the desired 1-sided confidence intervals can be obtained by substituting the appropriate upper and lower bound definitions from Box 2; primarily, this involves using 1-α levels of t-, z- or χ^2 statistics in place of 1-α/2 levels of the given two-sided formulae.

The general formulae in Box 2 can be used to compute two-sided 100×(1-α)% confidence intervals around sample estimates that are normally distributed (e.g., means, binomial proportions). Such confidence intervals are symmetric; i.e., the distance between the estimates and their upper or lower bounds, are equal. The distance between the estimate and its upper or lower bound is called the confidence interval half-width (W) and is a measure of the precision of the estimate; the smaller the distance the greater the precision. The size of the half-width depends on the size of the variance (S^2 for means and p(1-p) for proportions), the sample size and the specified confidence level. When each of the other two factors is held constant, the following changes will result in wider confidence intervals and thus less precise estimates of water quality:

1. increasing the variance
2. increasing the confidence level (e.g., going from 80% to 95% confidence)
3. decreasing the sample size

Specific formulae, and details for the construction of confidence intervals for a variety of population parameters are presented (with examples) in Appendix D.

C.1.6 Simple Random Sampling Designs

Having specified the tolerable error rates and the minimum sample size needed to assure them, the investigators can be reasonably confident that their resulting sample estimate(s) will meet their precision requirements. However, as Fig. 1a demonstrates, precision is not the only concern; the investigators must also insure that the sample estimates are unbiased (Fig. 1d). Assuming that sampling gear and laboratory procedures are not defective (i.e., negligible measurement error), bias can be controlled through proper implementation of an appropriate **sampling design**. Detailed treatment of this topic is available from the EPA document, *Guidance for Choosing a Sampling Design for Environmental Data Collection* (EPA QA/G5-S). In this section we present a brief overview of the sampling design process with an emphasis on the issue of representative sampling.

The first step in designing a sampling program is to construct the **sampling frame**. The sampling frame is simply a listing of all the sampling units in the target population. For example, consider a rectangular section of the benthos of a pond from which an estimate of the sediment concentration of pesticide X is desired. Assume further that the area has been divided into 160 equal-sized grid cells. Although the true population is an infinite number of points, the grid provides a convenient frame that completely covers the target area. Fig. 2a displays such a

grid with the center of each cell indicated with either an open or a solid circle. The grid cells are the population elements and hence the potential sampling units. A sampling frame is constructed by uniquely identifying each of the 160 sampling units; e.g., by numbering them from 1-160, starting in the upper left corner. Next, we select (without replacement) numbers in the range of 1-160 from a Table of random numbers to identify which population members to include in the sample. A set of n=30 randomly selected grid cells and a second randomly selected set of n=10 are denoted by solid circles in Figs 1a and 1b, respectively. The mathematical theory of combinations and permutations states that when n sampling units are randomly selected from a population of size N, it will be possible to draw S different, but equally likely samples:

$$S = \frac{N!}{n!(N-n)!} \tag{1}$$

Thus one could draw 2.749×10^{32} samples, each of size 30, or 2.274×10^{15} samples, each of size 10, from the 160 member target population. A simple random sample (SRS) is one in which each of the S possible samples has an equal probability (i.e., 1/S) of selection. For the benthic target population in Fig. 2, this insures that selection will not be biased for or against any part of the benthic area; i.e., the simple random samples are unbiased. However, closer inspection of Figs 2a and 2b reveals that, although unbiased, the two samples don't provide the same amount of information. Whereas sample 2a (n=30) provides information on pesticide concentrations from all four corners and the center of the target area, Sample 2b (n=10) lacks information from three of the four corners of the target area. Thus an SRS of n=30 appears to provide reasonably good coverage of the target population, while an SRS of n=10, does not.

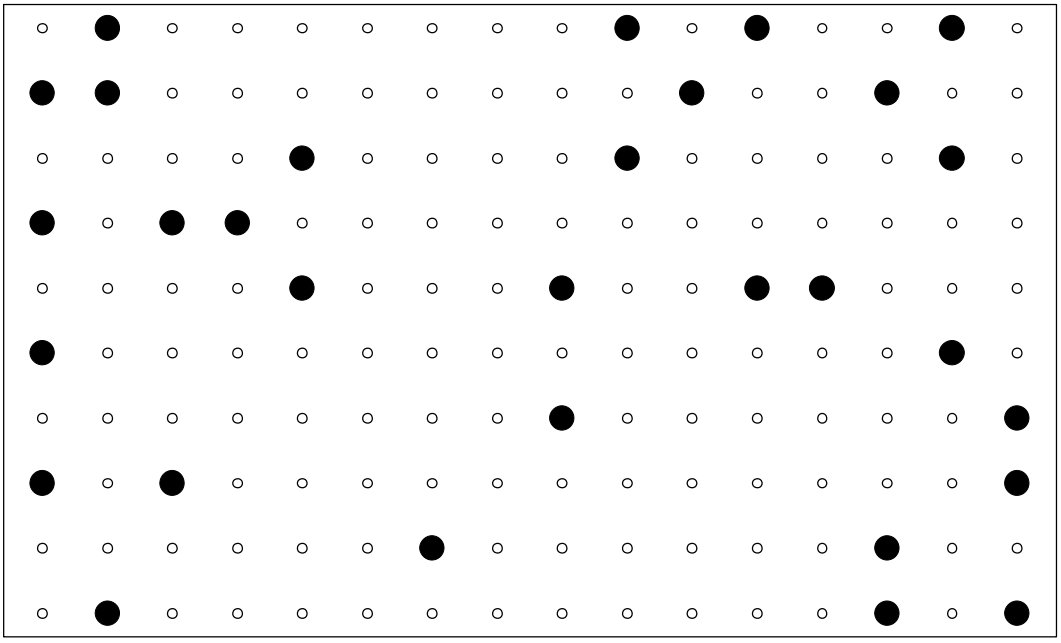
C.1.7 Representativness and Independence

The difference between the coverages of the two samples in Fig. 2 illustrates the concept of representativeness in the sampling of a target area. As illustrated, a **representative sample** is a sample, which, in microcosm, captures the range of the variability of the attribute of interest (e.g., the sediment concentration of pesticide X), among the elements of the target population. Note that unbiasedness in the sampling process does not necessarily insure representativeness of the resulting sample(s). While "representativeness" is relatively easy to conceptualize in a spatially referenced sampling frame such as Fig. 2, it is much more difficult to formulate an unambiguous mathematical definition.

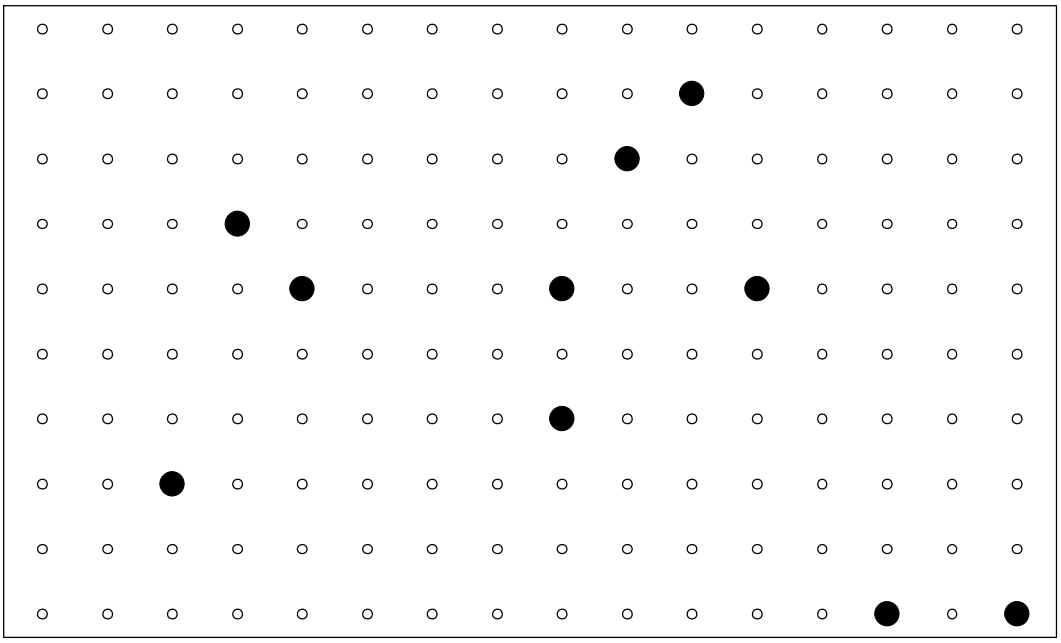
A sampling design that draws only from a specific subset of the target area, will yield unrepresentative samples. Unrepresentative sampling often occurs as a result of selectively sampling from the more accessible locations in the target population. For example, it may be much easier to obtain benthic samples from under bridges than from other stream reaches in a heavily forested area. To the extent that the measured characteristic (e.g., abundance of chironomid larvae) differs close to roadways vs. within the forest, the result will be an estimate like that shown in Fig. 1a or 1c. For this reason such **convenience samples** tend *not* to be representative. Similarly, samples taken close together in space and/or time tend to be more alike (i.e., **autocorrelated**) than more widely dispersed samples. For example, a sample of 30 sampling units from a single cove in a lake will not contain as much information as a sample of 30 sampling units, each coming from a different cove. This is because samples of autocorrelated

Fig. 2. Two examples of simple random sampling (SRS) from a 10x16 grid. Each circle represents a spatially fixed sampling unit; solid circles are sampling units that were randomly selected for inclusion to samples of sizes 30 (A) or 10 (B).

A. SRS with N=160 and n=30



B. SRS with N=160 and n=10



sampling units contain redundant information. To the extent that 30 autocorrelated sampling units are alike, they may carry no more information than 1 or 2 **independent** (i.e., uncorrelated) sampling units. Thus, for the purpose of making inferences about the variability of the measured attribute over the entire lake, the **effective sample size** of the sample from the single cove may be as small as one or two sampling units, compared to 30 from the multiple-cove sample.

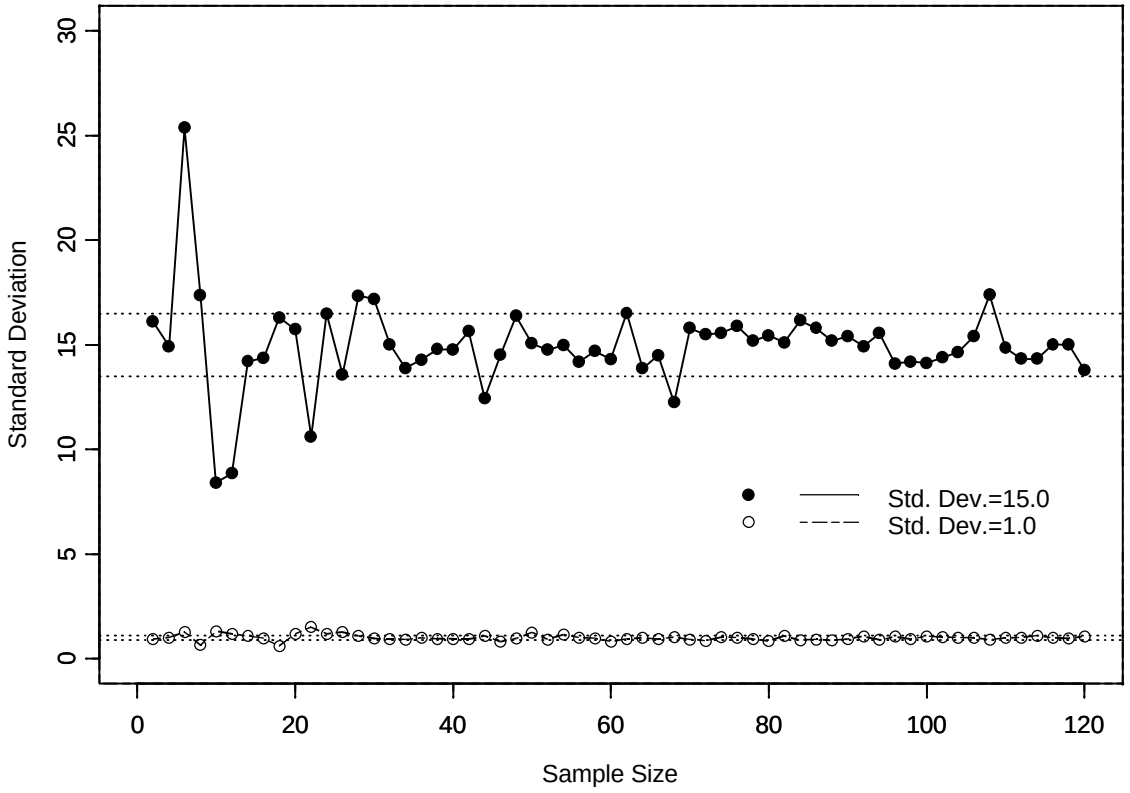
The problem of independence and sample size can be considered from variance component perspective. For example, a toxicologist may want to estimate the mean concentration of mercury in muscle tissue of a population of bass in a lake. Because of heterogeneity in the distribution of mercury exposure within the lake, heterogeneity of feeding behaviors, and physiological factors among individual bass, the tissue concentration of mercury can be expected to vary among fish within the population. If the sample mean is to be valid estimator of the mean of the population of bass in the lake, it must be computed from a representative sample of the bass target population. If the sample is truly representative, the variability (i.e., the heterogeneity) of the tissue concentration of mercury among bass in the sample will be an unbiased estimator of the mercury variability in the target population. Moreover, if the variability in the target population is high, the sample will need to be large (e.g., n=100 bass) in order to capture the variability of the population.

The among-bass variance in mercury concentration estimated from a probability sample of 100 fish selected from the lake population will be larger than a variance computed from 100 tissue samples taken from a single bass or from 100 bass taken from the same cove. In other words, neither the within-fish variability nor the within-cove variability will be the same as the among-fish variability for 100 different bass that were randomly selected from throughout the entire lake. In general, the heterogeneity of suprapopulations will be greater than that of component subpopulations because of the redundancies associated with correlations described above. Thus one should never limit sampling designs to subpopulations or subsets of the desired target population or target area.

The best way to ensure representativeness is to collect, using an unbiased selection method (e.g., SRS), a sufficient number of independent sampling units to capture the variability inherent in the target population. But how does one determine when the number of sampling units is "sufficient"? The best answer is, "conduct a pilot study". Using simulation techniques, Browne (1995) demonstrated that substituting the upper one-sided 1-α% confidence limit of the pilot sample standard deviation for S in Box 1a of Appendix D provided a sample size estimate sufficient to achieve the planned power (1-β), for a prespecified effect size in at least 1-α% of his Monte Carlo experiments.

Elliot (1977) described a more labor intensive (and more expensive) method in which a series of pilot samples with sample sizes increasing in increments of 5 sampling units is used to obtain a plot of the sample mean or variance against the sample size. The sample size at which the amplitude of fluctuation of the sample estimates damps indicates the minimum number of sampling units for a representative sample. This approach is illustrated in Fig. 3 in which sample sizes of single samples, increasing from n=2 to n=120, in increments of 2 sampling units, are plotted against the corresponding sample standard deviations. The open circles represent samples taken from a population with N=25,000, μ=20 and σ=1 while the solid circles represent

Fig. 3. Sample estimates of the population standard deviation plotted against the sample sizes of samples from two different populations. Open circles are estimates from a population whose true standard deviation is 1.0; solid circles are estimates from a population whose true standard deviation is 15.0. Broken lines enclose sample estimates that are within ± 10% of the true population value.



samples from a population with N=25,000, μ=20 and σ=15. Fluctuations in the standard deviations from the later population stabilize at about n=32, with sample estimates remaining within ± 10% of the population value (i.e., within the two broken lines), thereafter. Thus it appears that a sample of size n=32 is sufficient to represent the variance in the target population. Furthermore, increasing n from 32-120 does not appear to appreciably increase the representativness. By contrast, samples sizes as small as n=14 appear to quite representative of the population with σ=1. In general the less variable a population (i.e., the smaller σ), the smaller the required sample size. For example in an extreme case where every fish in a population of 100 is the same age, a sample of n=1 will be representative of the population age distribution.

C.1.8 Choosing a Sampling Design

The principle advantages of the SRS design are that (1) it provides statistically unbiased parameter estimates, (2) it is simple to understand and implement and (3) sample size calculation and subsequent statistical analyses of the sample data are straightforward. Simple random sampling is useful when the target population is relatively homogeneous; i.e., when there are no pronounced spatial or temporal patterns in the distribution of the population members or localized areas of extreme population abundance ("hot spots"). In those cases where SRS designs are not appropriate other probability-based sampling designs are available. Five of these, stratified sampling, systematic grid sampling, ranked set sampling, adaptive cluster sampling and composite sampling are reviewed in *Guidance for Choosing a Sampling Design for Environmental Data Collection* (EPA QA/G5-S). Each design has specific advantages in specific situations. Although they differ in details of selection of the sampling units, because they are all probability-based designs, they share two essential features: (1) each member of the target population has a known (though perhaps unequal) probability of selection into the sample; and (2) techniques of statistical inference (i.e., confidence intervals and hypothesis tests) can be applied to the resulting data. Data obtained from convenience or judgment sampling cannot be used to make formal statistical inferences unless one is willing to assume that they have the same desirable properties as probability samples, an assumption that usually cannot be justified (Peterson et al. 1999).

C.1.9 Data Quality Objectives Case History: Monitoring Dissolved Oxygen (DO) Downstream from an Agricultural Operation

This example concerns the use of a binomial proportion in planning for environmental decision-making. The example is presented in a continuous format to show the seven-step DQO Process in its entirety.

0. Background

Both flowing and standing waters contain dissolved oxygen (DO). Oxygen enters water principally from two sources: diffusion across the air-surface water interface and from photosynthesis of aquatic vascular plants and green algae. The capacity of water to hold oxygen depends on the complex interaction of several factors. Oxygen in water is negatively associated with temperature, altitude, salinity, plant and animal respiration, and decomposition of organic

matter by microorganisms. Generally dissolved oxygen decreases steadily from dusk to dawn and increases during the daylight hours.

Because oxygen is required for respiration, aquatic community structure, composition and health are to a great extent determined by the dissolved oxygen content of the waters they reside in. Warm-water fishes require a minimum dissolved oxygen concentration of 5.0 mg/l, while cold-water fishes require a minimum of 6.0 mg/l. Sustained exposure to dissolved oxygen concentrations < 2.0 mg/l for four days or more will kill most of the biota in an aquatic ecosystem. Such systems quickly become dominated by chemosynthetic bacteria (e.g., foul smelling sulfur bacteria) and undesirable air-breathing invertebrates.

Inputs of large amounts of organic material such as fertilizers or raw sewage encourage rapid proliferation of decomposing bacteria. Decomposition of large organic molecules (e.g., starches, fats, proteins, etc.) to smaller molecules (e.g., methane) is primarily an oxidative process that can rapidly deplete DO in the water column. Depletion may be transient, subsiding when the organic material has been completely decomposed or washed away or it may result in the permanent replacement of one aquatic community with another if the organic input continues for long periods of time. Thus dissolved oxygen is an important parameter that should be closely monitored, especially in ecosystems that are exposed to the risk of allocthanous inputs of organic materials from agricultural or industrial operations or from human residential development.

1. State the Problem

How were the planning team members selected? The planning team included a state limnologist (the project manager), a regional fisheries biologist, an environmental scientist from the EPA regional office, and a consulting statistician. The decision makers were the limnologist and the fisheries biologist.

How was the problem described and a conceptual model of the potential hazard developed? The problem was described as monitoring the dissolved oxygen concentration at a downstream location from a commercial hog operation. The operation has several large waste lagoons that have experienced overflow and runoff problems in the last 5 years. There have been several fish kills in the last 3 years just downstream of a large swale through which runoff from the hog facility has escaped in the past. The operators have recently made improvements to the lagoon system and the landscaping that are supposed to remedy the problem. The state pollution control agency has made a decision to closely monitor dissolved oxygen downstream of the swale for the next 3 years to determine if the problem has been resolved.

The conceptual model consists of a single discharge point into the river (i.e., the location of the swale), a half-mile mixing zone immediately downstream of the hog facility, and a cross-section of the river at a point approximately 0.60 of a mile from the swale and just beyond the mixing zone. The effluent consists of runoff of high organic material from the manure lagoons. Based on widely accepted models of stream dynamics it is assumed that mixing of the effluent plume with the ambient waters is homogenous at distances greater than 0.50 miles downstream of the discharge. At these distances microbial activity supported by the organic effluent may deplete dissolved oxygen to dangerous levels throughout the water column. Direct measures of

dissolved oxygen concentration (mg/l) taken along a cross-stream transect, located a half-mile downstream from the hog operation, were used to assess attainment of minimal DO standards.

What were the available resources and relevant deadlines? The state pollution control authority was prepared to commit a maximum of $5000 per year, for a 3-year period, to monitor dissolved oxygen concentrations in the vicinity of the hog operation. Monitoring was scheduled to begin in January 2000.

2. Identify the Decision

What was the Decision Statement? Consistent with existing state water quality standards, the decision statement was to determine if more than 10% of the DO estimates from regularly collected downstream samples fell below the 5.0 mg/l standard in any full calendar year of monitoring.

What were the alternative actions? If statistically significant nonattainment of the 10% DO standard was observed in any year during the 3-year monitoring period, the water would be listed and appropriate actions would be taken against the operators of the hog facility.

3. Identify the Inputs to the Decision

Identify the kind of information. Assessment of DO was made by direct *in situ* measurement of DO in a cross-section of the river approximately 0.60 miles downstream of the hog facility.

Identify the source of information. The state pollution control authority states that no more than 10% of the DO samples taken from a reach of river in a single year may be less than 5.0 mg/l.

What sampling and analytical methods were appropriate? DO measurements were made with a YSI Series 5700 DO probe, following procedures described in section 6 of the USGS National Field Manual for collection of water-quality data. Preparation, maintenance and calibration for the YSI 5700 were carried out following specifications at:

http://water.wr.usgs.gov/pnsp/pest.rep/sw-t.html

Dissolved Oxygen readings were recorded in the field. The results provided information on DO concentration at the monitoring site on each sampling date. The probe detection limit was well below the 5.0 mg/l action level.

4. Define the Boundaries of the Study

What population was sampled? Water passing through a cross-section of the river located 0.60 miles downstream from the hog facility.

What were the spatial boundaries? The spatial boundaries of the monitoring site were defined by a transect dawn perpendicular to the river, between two permanent markers each located 0.60 miles from the commercial hog operation. The transect spanned the width of the

river and was 75 meters across. The depth along the transect between 5 and 70 meters from the left bank ranged from 2.6 to 3.7 meters with a mean of 3.1 meters.

What was an appropriate time frame for sampling? Sampling was scheduled to commence in January of 2000 and run through December of 2002.

What were the practical constraints for collecting data? There were no specific practical constraints in collecting the data within the specified time frame at the downstream sampling station.

What was the scale of the decision-making? The area-adjusted mean DO computed from DO measurements taken along the transect on a specific sampling date was the basic unit of analysis. The area-adjusted mean represents the DO in the cross-section of the stream lying beneath the transect on a specific sampling date.

5. Develop a Decision Rule

What was the decision rule and Action Level? Because of its importance to aquatic community structure and health, state DO criteria permit no more that 10% of the samples taken from a body of water during a year the have less than the 5.0 mg/l minimum. This proportion was computed by comparing each area-averaged mean DO to the standard and dividing the number of nonattainment values by the total number of sampling times in a year. The 10% criterion is the action level. If the true proportion of nonattainments exceeded the action level, remedial action and listing of the river was indicated.

6. Specify Tolerable Limits on Decision Errors

How was the baseline condition set? The baseline condition assumed by the investigators was that the water in the cross-section attained the DO criterion is each year in which it was monitored. This decision was made because the baseline is traditionally assumed to be attainment. However in this case, the choice of the baseline was trivial because the investigators intended to specify balanced false-acceptance and false-rejection error rates.

How was the gray region specified? The gray region was designated by considering the consequences of distinguishing between 10% and 25% nonattainment. The benefits of distinguishing among relatively small exceedance rates within the gray region were deemed insufficient to justify the cost (Table 1). Thus, following Smith et al. (2000) the upper bound of the gray region was set at 25% nonattainment.

How were tolerable decision error limits set? The consequences of not detecting low DO events need to be balanced with the consequences of falsely declaring that the reach of the river immediately downstream of the hog operation had seriously depleted dissolved oxygen. The consequences of the former include fish-kills and undesirable community restructuring. The consequences of the latter are unnecessary economic hardship on the operators of the commercial hog facility and perhaps incorrectly listing the river. In this case, it was decided that the two types of error were equally undesirable. Thus, a decision was made to simultaneously

minimize the false negative and false positive error rates to the same value. Given the fiscal constraints on the monitoring program, it was determined that the common error rate associated with a gray region of width 0.15 should be no greater than 0.15. This is illustrated in Figure 4 and Table 2.

7. Optimize the Design for Obtaining Data

What was the selected sampling design? Following the recommendations in Section 6.0.2B of the USGS National Field Manual for collection of water quality data, DO measurements were made *in situ* along a fixed transect using the Equal-Width Increment method (EWI). The method calls for the transect to be divided into at least 15 increments of equal width. DO measurements were made in the middle of each interval of the transect at the middepth of the vertical between the surface and the stream bottom. The surface-bottom depth was recorded along with the DO in each interval. The cross-sectional area of each increment was computed as the product of the surface-bottom depth and width of the interval. The total cross-sectional area of the stream at the sampling station was computed as the sum of the areas of the all of the intervals. Finally, the area-adjusted mean DO of the cross-sections of the river at the sampling station was computed as:

$$\overline{DO} = \sum_{i=1}^{n} w_i DO_i \qquad (1.1)$$

where w_i = the cross-sectional area of the i^{th} interval divided by the total cross-sectional area of the stream at the sampling station

DO_i = the middepth DO concentration at the center of the i^{th} interval of the transect

n = the total number of intervals that the transect was partitioned into.

The distance of the transect from shore-to-shore at the sampling station was 75 meters. The transect was divided into 15 increments, each of 15 meters in width. Based on the cost constraints, the width of the gray region, and the 15% balanced error rates, it was determined that DO should be measured at the sampling station at 11-day intervals beginning on January 3 of each year and ending on December 31, for a total of 34 evaluations per year (@ $150 per evaluation), for an annual cost of $5100. It was decided that the budget should be adjusted to permit the small ($100) cost overrun.

Fig. 4. Decision performance goal diagram for a Z-test based on the normal approximation to the binomial distribution. H_0: population proportion of DO concentrations that is ≤ 5.0 mg/l, is ≤ 0.10 vs. H_a: population proportion > 0.10, when the sample size=33 and the width of the gray region is 0.15 (minimum detectable effect size=d=0.15).

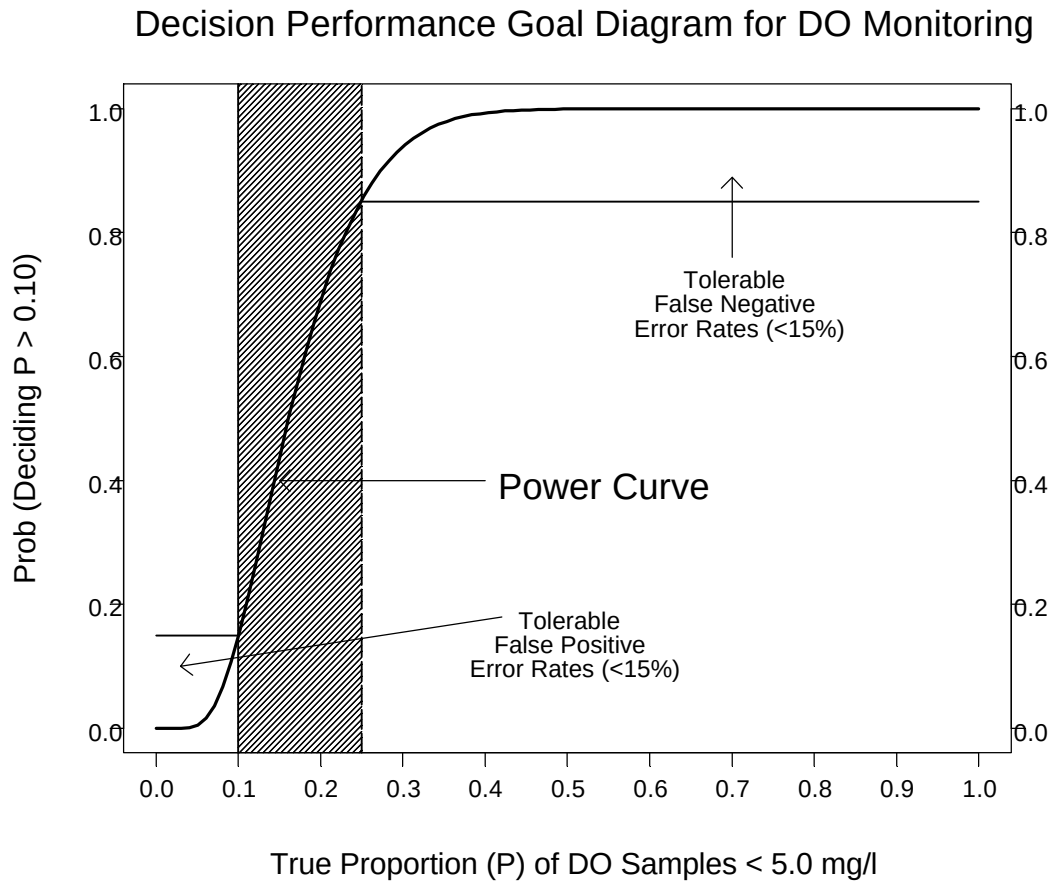


TABLE 1. EFFECTS OF INCREASING THE WIDTH OF THE GRAY REGION ON SAMPLE SIZE AND COST

WIDTH OF GRAY REGION	FALSE POSITIVE RATE	FALSE NEGATIVE RATE	SAMPLE SIZE	COST
0.01	0.15	0.1500	4036	$605,400
0.05	0.15	0.1494	186	$27,900
0.10	0.15	0.1485	53	$7,950
0.15	0.15	0.1473	26	$3,900

TABLE 2. EFFECTS ON THE SAMPLE SIZE AND COST OF INCREASING BALANCED ERROR RATES

FALSE POSITIVE RATE	FALSE NEGATIVE RATE	SAMPLE SIZE	COST
0.05	0.0481	65	$9,750
0.10	0.0983	40	$6,000
0.15	0.1473	26	$3,900

What were the key assumptions supporting the selected design? Four assumptions were required to support the sampling design:

1. Flow velocity was relatively uniform along the transect
2. Stream depth was relatively uniform along the transect
3. There was homogenous mixing of any effluent from the hog operation with the ambient waters between the discharge point and the sampling station
4. The 11-day sampling interval was sufficiently small that no significant low-DO events would be missed

Assumptions 1 and 2 were verified by direct measurement along the transect prior to initiation of the study. Assumption 3 was accepted based on numerous published limnological studies of mixing dynamics in rivers. The 4th assumption represents a compromise between the desire for good coverage of the time period of concern and the project cost constraints. Although short-term transient episodes of oxygen depletion might be missed with an 11-day sampling interval, it was decided that the cost of more frequent sampling could not be justified given that short-term depletions severe enough to cause significant fish-kills would be observed and reported in any event. A systematic sample of the kind used in this study was deemed sufficient to capture patterns of "typical" fluctuations in DO during a 3-year period at the monitoring site.

C.2 The Data Quality Assessment Process: Exploratory Data Analysis

C.2.0 Review of the Steps in a Basic DQA

The data quality assessment (DQA) process is described in detail in the EPA document, *Guidance for Data Quality Assessment EPA QA/G-9* (EPA/600/R-96/084). EPA's DQA process is a 5-step plan designed to provide scientifically defensible conclusions through statistical analyses of environmental data previously collected in accordance with a well-conceived DQO strategy. The DQA steps are:

1. Review the study DQOs and associated sampling design(s)
2. Conduct a preliminary review of the data
3. Based on the data and the research question, select an appropriate statistical test(s)
4. Verify the underlying assumptions of the selected test(s)
5. Draw conclusions from the data analyses

Activities in DQA Step 1 should focus on:

(a) Review of the outputs of the four statistical DQOs (i.e., target population definition, decision rule(s), acceptable limits of decision error, and sampling design)
(b) Translation of the decisions identified in the DQOs into statistical hypotheses
(c) Confirmation of the limits on the decision error rates
(d) Any features of the sampling design that would bear on the selection or interpretation of the statistical test(s)

The remainder of this Appendix will focus on techniques and considerations that are useful for carrying out steps 2-5 of the DQA process. We begin with a discussion of a graphically based approach to Step 2 which coincidentally contributes to the resolution of many of the issues associated with the subsequent steps in the DQA process. In Sections C.3.0 – C.3.3 we will review the basic principles of statistical hypothesis testing with emphasis on those aspects that bear on control of decision error rates. Various types of statistical tests appropriate for water quality assessments will be reviewed and some examples of their application and interpretation will be presented in Appendix D.

C.2.1 Exploratory Data Analysis: Basic Principles

The second DQA step involves an initial quality assurance (QA) review of the data for data entry errors, etc. This is followed by computation of the statistics specified in the DQO process and/or any auxiliary statistics that the analyst might deem necessary to the interpretation of the data or to the assessment of the assumptions of the statistical tests. Finally, and perhaps most importantly, step 2 requires extensive **exploratory data analyses (EDA)**. EDA usually involves graphical methods that facilitate identification of characteristics and/or relationships in the data that are often crucial to proper interpretation of the subsequent statistical tests. Indeed, the EDA results may suggest that additional or different tests should be applied (Cleveland 1993), or that some transformation of the data may be necessary. Graphical EDA methods are described in detail in *Guidance for Data Quality Assessment EPA QA/G-9* (EPA/600/R-96/084), *Biological*

Criteria: Technical guidance for Survey Design and Statistical Evaluation of Biosurvey Data (EPA 822-B-97-002*)*, and in *Visualizing Data* (Cleveland 1993). These techniques are easily implemented with standard commercial statistical software (e.g., SASGRAPH, SPLUS, SPLUS EnvironmentalStats and SPLUS Spatial).

It was pointed out in Section C.1.5 that some confidence interval estimators require that the sampling units and/or the sampling distributions be approximately normally distributed. Similar assumptions of normality will be required for the validity of many of the statistical tests in Appendix D. In Section C.1.7, additional requirements for spatial and temporal independence among the sampling units were imposed for the estimation of confidence intervals. The independence requirements also apply to the statistical hypothesis tests that are discussed in Appendix D. There are several ways to assess the validity of these assumptions but the most effective methods involve graphical EDA procedures (Cleveland 1993). Examples of graphical verification of normality and of graphical assessment of spatial and of temporal independence will be illustrated in this section of the appendix.

C.2.2 EDA Example 1: Assessing Normality of Continuous Data

Two hundred and forty-four turbidity readings (Table 3) were recorded from 1980-2000 in a reach of the Mermentau River in Southwest Louisiana. From the sample, the investigators computed the following descriptive statistics:

n=244
Minimum = 8
Median = 75.5
Maximum =600
Mean=100.3
Std. Deviation =86.1

The investigators wanted to determine if the mean turbidity was greater than the local criterion value of 150 NTU. Two options were available to them: (1) they could compute an upper 1-sided 95% confidence interval (Appendix D, Box 3) on the mean turbidity and check to see if 150 was included in within the interval or (2) they could compute a 1-sided t-test (Appendix D, Box 8). Both methods require that the distribution of the 244 turbidity readings be approximately normal. The most important attributes of the normal distribution are that it is symmetric and that its mean and median are equal. A quick examination of the turbidity descriptive statistics reveals that these conditions do not hold for the sample, suggesting that the turbidity data are not normal.

The best way to assess the form of a population distribution is to graph its frequency distribution. A frequency distribution is a histogram that displays the way in which the frequencies (i.e., counts) of members of the sample are distributed among the values that they take on. The corresponding **relative frequency distribution** (Fig. 5A) can be calculated by dividing each count by the total sample size. The familiar "bell curve" is a graph of the expected relative frequency distribution of a variable that is normally distributed (e.g., heights or weights of men or women). When the relative frequency is based on smaller sample sizes, it will take the form

of a histogram composed of clearly discernible individual bars; when it is computed from very large numbers of individuals (e.g., hundreds of thousands of women) its graph tends to smooth out as the bars increase in number and merge together to produce a smooth surface (i.e., the bell).

If the counts of the individuals are large enough, it is often possible to summarize the relative frequency distribution with a mathematical expression called the **probability density function** (PDF). The PDF predicts the relative frequency as a function of the values of the random variable and one or more constraining variables, called model parameters, which can be estimated from the sample data. Continuous distributions whose PDFs can be so defined are called **parametric continuous distributions**. The PDF is the algebraic expression for the line that delimits the shape of the relative frequency distribution. In Fig. 5B, the plot of a PDF for a normal distribution (bell-curve) is superimposed on the relative frequency distribution of the sample of turbidity values (Table 3) from which it was computed. The specific form of the bell curve is controlled by the 2 parameters of the normal distribution: the population mean (μ) and the population standard deviation (σ). The larger the mean, the father the center of the bell is shifted to the right; the larger the standard deviation, the wider and lower the bell.

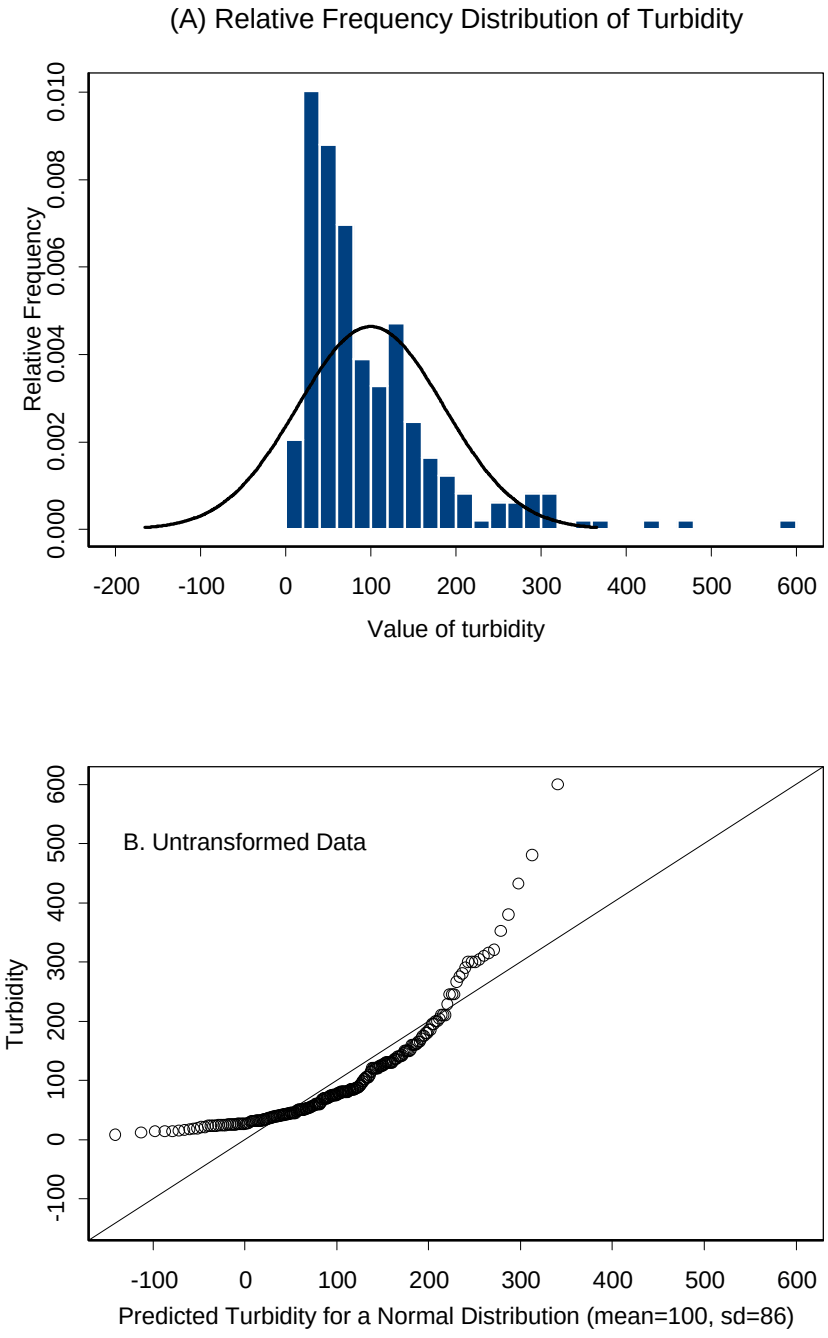
In Fig. 5A, the turbidity data have been divided into 21 groups, each of which has a vertical bar associated with it. The height of the bar indicates the proportion of the 244 turbidity measures that are in the group, which in turn, is an estimate of the probability of any turbidity measure taken from the river reach being in the group. The bell curve is based on the assumption that the actual data come from an underlying normally distributed population with μ= the sample mean=100.3 and σ =the sample standard deviation=86.1. It is clear from Fig. 5A that the observed relative frequency distribution does not fit the assumed normal distribution very well. Specifically, the normal distribution predicts a substantial number of negative turbidity values, far fewer turbidity values in the range of 20-80 NTU than were actually observed and no values > 360.

Although Fig. 5A employs techniques (i.e., histograms and bell-curves) that are familiar to most scientists, it has the drawbacks that it requires significant effort to produce and that subtle differences between the assumed normal and the sample distributions can be obscured by the user's choice of the number of groups into which to divide the data (21, in this case). Figure 5B, called a Q-Q-plot or a normal probability plot, is a much simpler graph that can be quickly and accurately read and interpreted. The Q-Q-plot is a graph of the sample data values (vertical axis) against the values predicted by the normal probability curve (horizontal axis), with μ and σ equal to the mean and standard deviation of the sample data. If the sample data exactly matched the normal predicted values, all the data points would lie on the diagonal line. In Fig. 5B, data points corresponding to the upper and lower tails of the sample distribution clearly deviate from normality. Because the deviant points are above the diagonal, it is clear that the tail values are considerably larger than would be expected for a normal distribution. Furthermore, it is easy to see that the turbidity data in the center of the distribution (i.e., 80-260 NTU) have values that are close to normal, but are slightly smaller than expected (i.e., they are located just below the diagonal).

Table 3. Monthly turbidity values (1980-2000) for a reach of the Mermentau River in Southwest Louisiana. Data are sorted by NTU (Nephelometric Turbidity Unit) value. Note that 232/244 sample values are less than 300 NTU; thus there are only 12 values between 300 and the maximum value of 600 NTU, indicating a long-tailed, highly skewed distribution.

8	12	14	14	14
15	16	17	18	19
21	21	23	23	23
23	24	24	24	24
25	25	25	25	25
26	26	26	26	26
27	27	28	30	31
31	31	32	32	32
32	32	33	34	34
34	35	36	36	37
37	38	38	39	39
39	40	40	40	41
41	41	42	42	43
44	44	44	45	45
45	45	45	45	48
48	50	50	50	50
50	50	50	52	52
53	53	53	54	54
54	56	56	57	57
58	60	60	60	60
60	60	62	64	68
68	70	70	70	**70**
70	71	72	72	74
74	74	75	75	75
75	**75**	**76**	78	78
78	78	80	80	80
80	80	80	80	80
80	81	82	83	85
85	85	85	85	86
87	88	88	88	90
91	93	95	98	98
102	102	105	105	105
105	108	114	115	120
120	120	120	120	120
120	123	123	125	125
125	126	128	128	130
130	130	130	130	130
130	133	136	136	136
140	140	140	140	142
145	150	150	150	150
150	152	160	160	160
160	162	165	165	170
175	175	180	180	185
185	195	195	200	200
204	210	210	210	228
245	245	245	266	275
280	290	300	300	300
304	310	315	320	352
380	432	480	600	

Fig. 5. Distribution of 244 monthly turbidity measurements (NTU) from the Mermantau River, 1980-2000. (A) Relative frequency histogram and normal probability function (μ = 100.3, σ = 86.1). (B) Normal Q-Q plot comparing distribution of the turbidity data to their expected values (diagonal) under the assumption that they are normally distributed with μ = 100.3 and σ = 86.1.



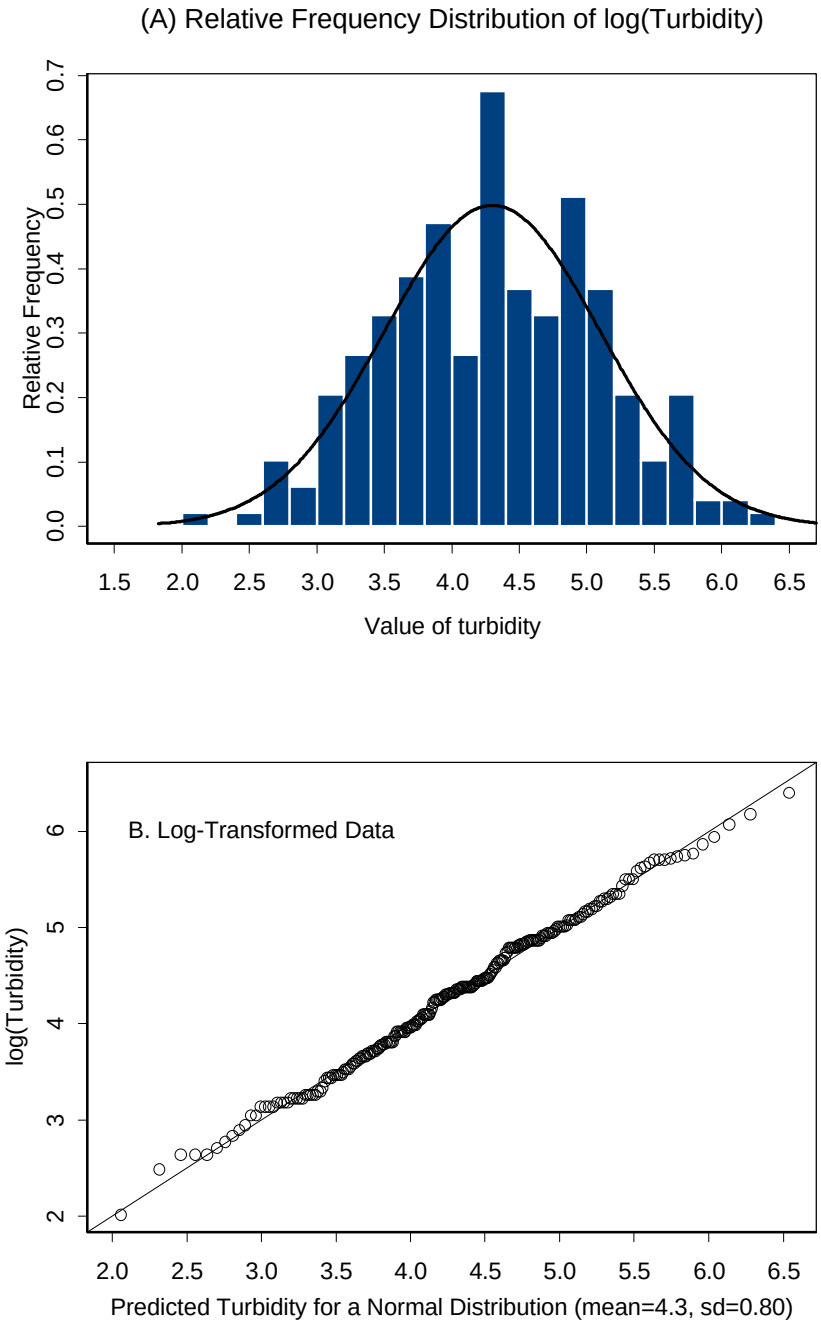
Normal probability plots are available from most commercial statistical software and can be produced with a single mouse click or a line or two of simple programming code. Alternatively, a number of formal statistical tests for the normality assumption are available from the same software packages. However, in practice, these tests (called goodness-of-fit tests) are not as useful as the graphical EDA methods. As with all such tests, they generate p-values that tell the user very little about the actual shape of the data distribution. Moreover, the GOF-test results are strongly dependent on the sample size. The normal GOF tests tend to accept the normality assumption for small sample sizes and reject it for large sample sizes, irrespective of the actual shape of the distribution of the sample data. The reasons for this troublesome behavior have to do with sample size and power relationships that are explained in Section C.3.2.

Having determined from the Q-Q plots and the sample descriptive statistics that the turbidity data are not normal, two courses of action are available: (1) find another test-statistic or confidence interval estimator which does not require normality or (2) apply a suitable transformation to the data such that the transformed values will be approximately normally distributed. Discussion of the first alternative will be taken up in section C.2.3. Here we will demonstrate how to apply the log-transformation and evaluate its effectiveness.

As noted, the distribution of the turbidity data is concentrated in the range of 0-200 with a very long tail of large values extending out to 600 NTU. This type of left-skewed, asymmetric distribution is very commonly observed for environmental and ecological variables. Influenced by the large values in the right tail, the sample mean of such a distribution will usually be substantially larger than the sample median (e.g., 100.3 vs. 75.5 for the turbidity data). A left-skewed distribution of positive-valued environmental data with mean > median is usually log-normally distributed (i.e., the distribution of logarithms of the raw data values is normally distributed). Fig. 6A displays the distribution of the logs of the turbidity data, overlaid with a plot of a bell-curve with μ= the log-scale sample mean=4.30 and σ =the log-scale sample standard deviation=0.80. The corresponding Q-Q-plot confirms that the turbidity data are approximately lognormally distributed.

There is an obvious problem in using log-transformed data for water quality attainment decisions; i.e., water quality standards are measured on the original scale. However, because of some fortuitous relationships among the means and medians of normal and lognormal data, this does not present a serious difficulty. Statistical inference based on the means of normal data depends on the fact that the mean and the median of normally distributed data are equal. Thus when a difference of say k-units is demonstrated between the mean of a normal distribution and a criterion value, one can assume that the center of the distribution also differs from the criterion by k-units. The back-transform of the log-scale mean [e.g., exp(4.3) for the turbidity data] of a log-normally distributed variable is called the geometric mean and is a close approximation to the median of the data on the original untransformed scale. Therefore, inferences based on the geometric mean of a lognormal distribution are equivalent to inferences based on the mean of a normally distributed variable. Comparison of the geometric mean to the criterion value should be interpreted no differently than a comparison involving the mean of a normal distribution.

Fig. 6. Distribution of 244 monthly log-transformed turbidity measurements from the Mermantau River, 1980-2000. (A) Relative frequency histogram and normal probability function ($\mu = 4.3$, $\sigma = 0.80$). (B) Normal Q-Q plot comparing distribution of the log-turbidity data to their expected values (diagonal) under the assumption that they are normally distributed with $\mu = 4.3$ and $\sigma = 0.80$.



The equivalence of the geometric mean and the median of the turbidity data can be easily demonstrated. The nonparametric estimate of the median turbidity and its 90% confidence limits (see Section D.1) is:

75.5(70.0, 80.0)

Noting that the Z associated with a 90% 2-sided confidence interval is 1.645 and that the log-scale mean and standard deviation are 4.3 and 0.80, we can apply the first equation in Box 1 of Appendix D (substituting z for t) to form a 2-sided 90% confidence interval on the log-scale mean:

4.301 (4.217, 4.385)

To obtain the geometric mean and its 90% confidence interval we simply exponentiate each of the values in the preceding expression:

73.8 (67.8, 80.2)

The geometric mean and its confidence interval differ only very slightly from the median estimate. The small disparity this is due to the fact that the distribution of the turbidity values is only approximately lognormal (Fig. 6).

C.2.3 EDA Example 2: Assessing Normality of Count Data

We have seen that the log-transform is appropriate for normalizing data for a skewed **continuous random variable**. However, it is also common to encounter distribution of skewed **discrete random variables** in environmental and ecological sampling; e.g., the distribution of snail counts among sampling units taken from the littoral zone or of exceedances in a collection of 1-liter sampling units from a lake. Very often the frequency distribution of such counts will approximate a Poisson distribution. Samples from a Poisson population are easily recognized by the fact that their mean and variance are equal or nearly so. Statistical theory insures that the distribution of the square roots of Poisson-distributed data is normal. Thus, the square-root transformation is recommended for normalizing skewed count data.

Roadside counts of meadowlarks from a survey of roads in the agricultural region of Southwestern MN in 1990 are presented in Table 4. The survey was designed so that an observer, driven over ninety different 5-mile stretches of road in the target area, counted the number of meadowlarks that were visible from the vehicle. In this design, each 5-mile stretch is a sampling unit, and the individual counts are the responses measured on them. The data were used as part of a long-term monitoring program to assess trends in the mean of the counts. The following descriptive statistics were computed from the 1990 sample data:

n=90
Minimum = 2
Median = 38
Maximum =143
Mean=46.1
Std. Deviation =36.2

Table 4 Meadowlark Counts

2	3	4	4	4
5	6	7	8	9
10	10	10	10	12
12	12	14	14	14
16	16	17	17	19
19	20	21	21	22
22	22	23	23	25
27	29	30	30	31
32	35	36	37	38
38	**38**	40	41	41
42	45	45	47	48
50	51	51	52	53
54	55	56	57	57
58	59	62	72	73
73	74	79	82	89
92	95	95	98	99
102	103	107	108	109
111	128	135	141	143

The relative frequency distribution of the counts, and an overlaid normal curve for a population with $\mu = 46.1$ and $\sigma=36.2$ are plotted in Fig. 7A. As suggested by the fact that the sample mean (46) is nearly 28% larger than the median, the distribution of counts in the sample is skewed. Similar to turbidity data, the Q-Q plot of the meadowlark counts (Fig. 7B) demonstrates that the most serious departures from normality occur in the tails of the distribution. However, the upper tail of the bird-count distribution, though longer than expected for a normal distribution, is much shorter than the upper tail of a lognormal distribution (compare to Fig. 5A). This is typical of Poisson data.

The results of the square-root transformation are displayed in Figs 8A and 8B. Although some lack-of-fit is still evident in the tails of distribution, the Q-Q plot indicates that the overall fit of the transformed counts to the normal distribution is quite good. As was the case with the log-transformed data, the back-transform of the mean of the square roots (i.e., 6.2^2) provides a close approximation to the median of the original data (i.e., 38.0). However well the square-root transform seems to have worked in this case, it is noteworthy that the mean and variance of the untransformed data are far from equal. This would seem to contradict the earlier statement that equality of the mean and the variance was an indicator of "Poissonness" and a justification for employing the square-root transformation to normalize the data.

Actually, this example illustrates the robustness of the square-root transformation. The major benefit of the log and square-root transformation is that they lead to an estimate of the median of the original distribution. But, it turns out that mean and median of almost any symmetric distribution will be equal or nearly so. Thus a transformation applied to a skewed distribution does not have to normalize it, it only needs to make it roughly symmetric. This is a less demanding requirement than normalization, hence the square-root and log transformations tend to work quite well on many skewed distributions even if they are not quite Poisson or lognormal. So, in practice, it is worth applying one or the other of the two transforms to all skewed data as a matter of routine, followed by verification of the results with Q-Q plots. If one of these transformations does not correct the skew, one of the nonparametric methods described in Section D.3 should be considered.

C.2.4 EDA Example 3: Assessing Spatial Independence

The concept of statistical independence among sampling units and its importance were introduced in Section C.1.7 of this appendix. Spatial *and* temporal independence are required for all of the inferential procedures described in this appendix. The following example illustrates the nature of the problem of spatial autocorrelation and its diagnosis; Example 4 will illustrate graphical analysis of temporal autocorrelation. Although counts of organisms are analyzed in the following example, the graphical diagnostics for the effects of spatial autocorrelation are equally applicable to spatially correlated distributions of chemical (e.g., pesticide residue concentrations) and/or physical variables (e.g., Secchi depths) in a target water. Moreover, the issues discussed here apply equally to sampling designs based on grids, transects, riffles or any other sampling units or clusters.

Fig. 7. Distribution of 90 meadowlark counts from SW Minnesota, 1990. (A) Relative frequency histogram and normal probability function ($\mu = 46.1$, $\sigma = 36.2$). (B) Normal Q-Q plot comparing distribution of the count data to their expected values (diagonal) under the assumption that they are normally distributed with $\mu = 46.1$ and $\sigma = 36.2$.

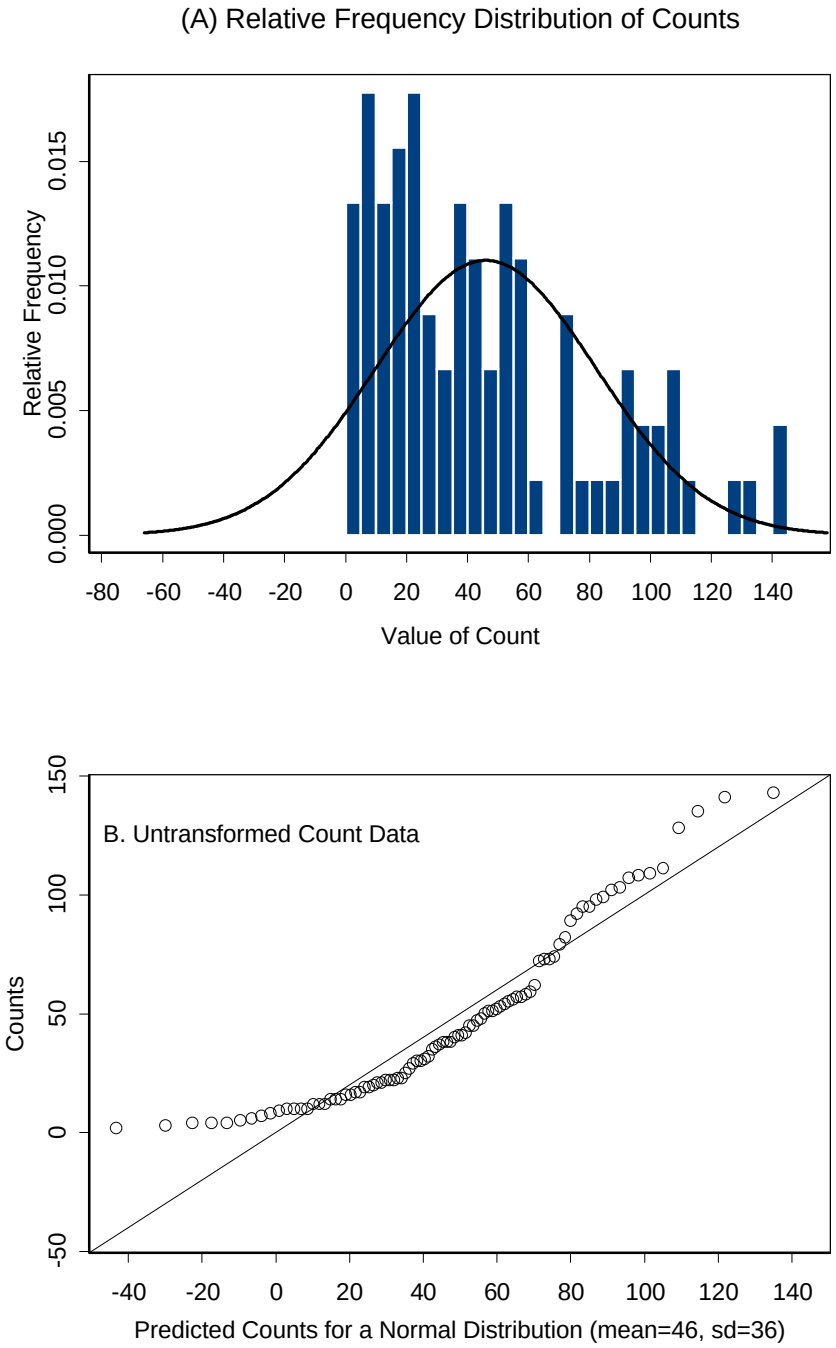
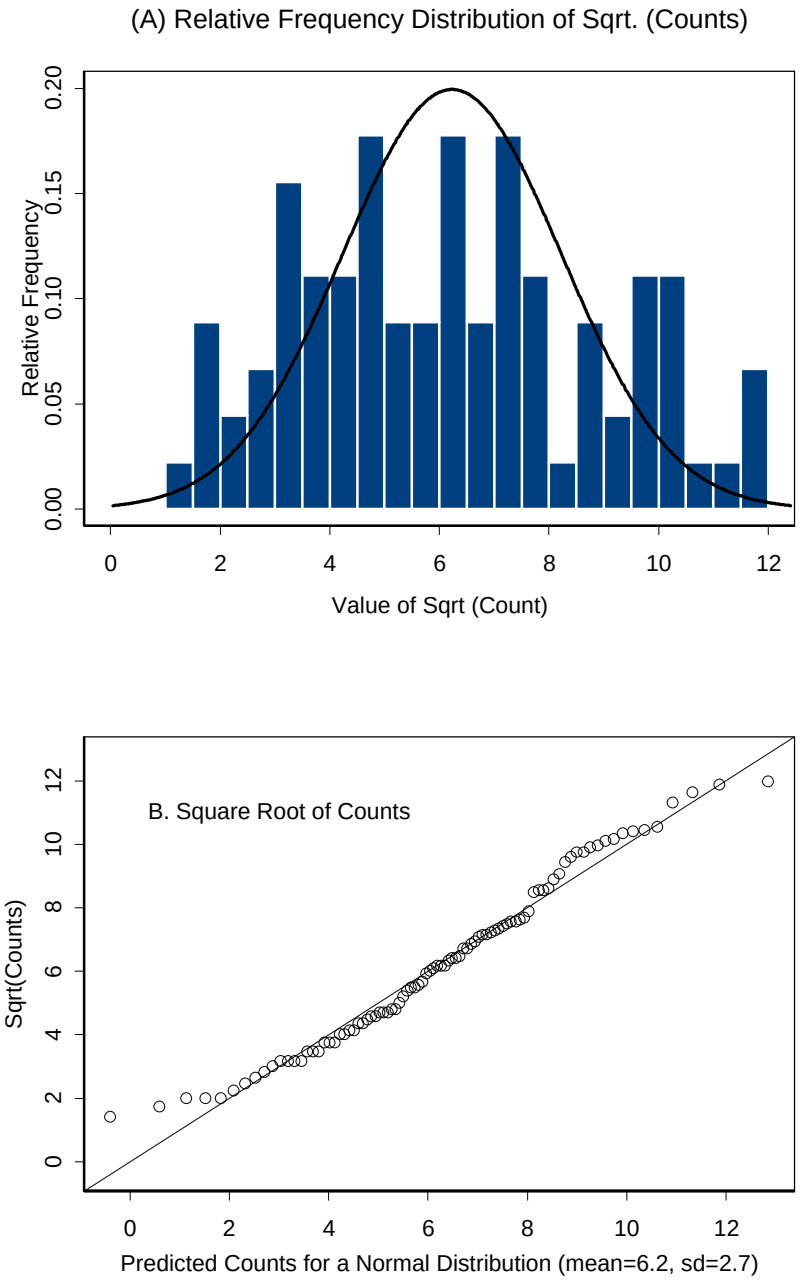


Fig. 8. Distribution of square roots of 90 meadowlark counts from SW Minnesota, 1990. (A) Relative frequency histogram and normal probability function ($\mu = 6.2$, $\sigma = 2.7$). (B) Normal Q-Q plot comparing distribution of the square roots to their expected values (diagonal) under the assumption that they are normally distributed with $\mu = 6.2$ and $\sigma = 2.7$.



Hengeveld (1979) conducted a study on the ecology of two species of beetle in a coastal flood plain. The study site was grided into a 21×12 rectangle of square cells, each of which was 40m on a side. Counts of each species were made once in the spring and once in the fall of 1975. In this design, each of the 252 grid cells is a sampling unit and the individual beetle counts are the measured responses. For this example we will only consider the spring counts of one of the species, *Dychirius globosus*. Basic descriptive statistics for the spring sample are presented below.

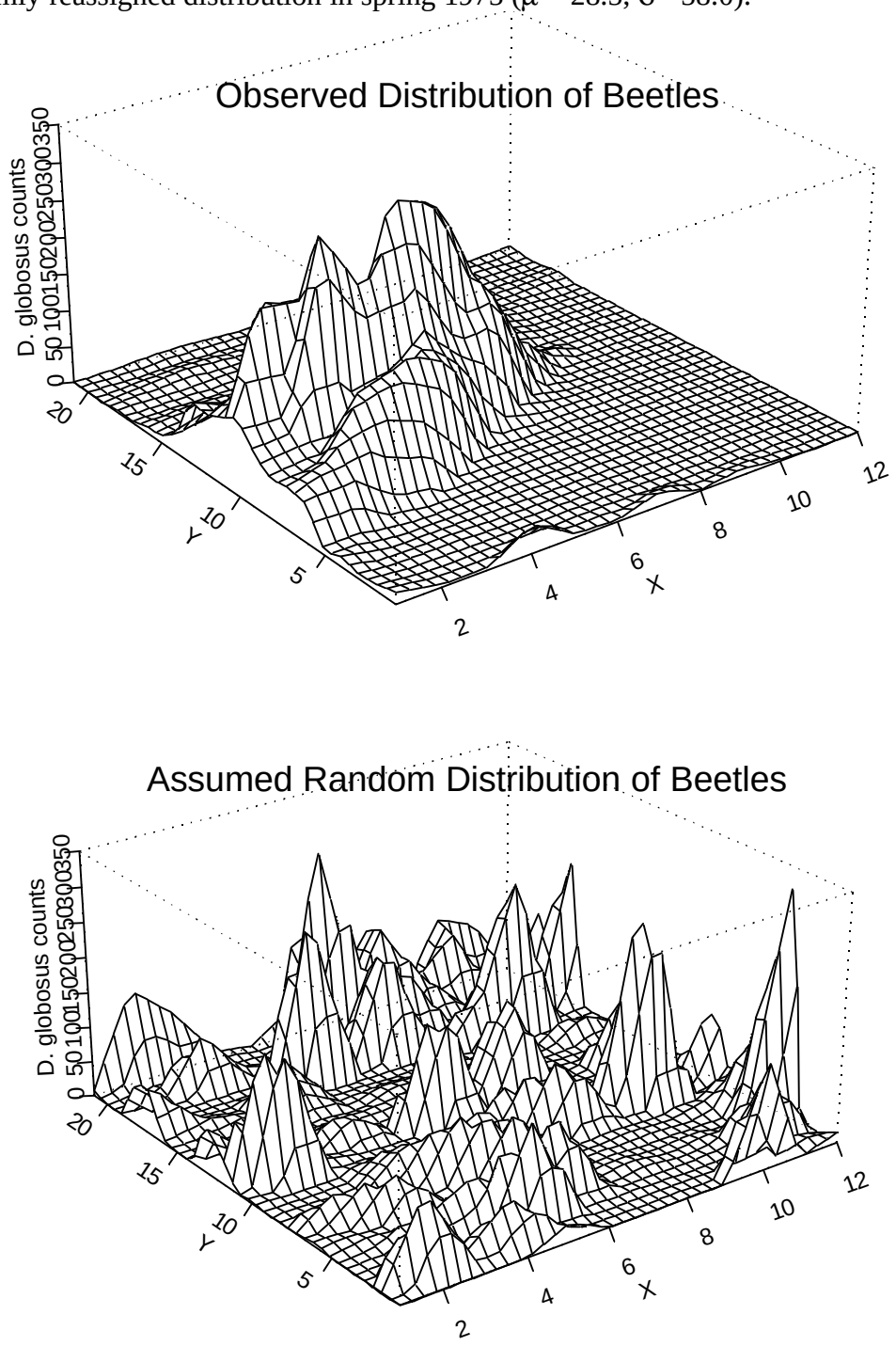
n=252 grid cells
Minimum = 0 beetles
Median = 2 beetles
Maximum =334 beetles
Mean=28.5 beetles
Std. Deviation =58.0

Given the large disparity between the mean and the median and between the mean and the standard deviation, we can conclude that the data are severely skewed and thus would require some sort of normalizing transformation before any of the parametric statistical tests described in Appendix D could be applied. However, all of these inferential procedures also require the data to be spatially and temporally independent. The independence assumptions cannot be assessed by inspection of sample descriptive statistics of the sort shown above.

Before considering the types of statistics that are appropriate for the evaluation of spatial autocorrelation, the nature of the problem needs to be illustrated. A 3-dimensional plot of the spatial distribution of the beetle counts on the sampling grid is shown in Figure 9A. The X- and Y-axes are the latitudinal and longitudinal dimensions of the sampling grid and are divided into grid cell units. The third axis provides a 3-dimensional surface representing the beetle counts. The graph reveals an area of intense population density near the center of the grid, with a complete absence of beetles elsewhere. In fact, 81 of the gird cells have counts of zero. Thus, if the beetle count in any given grid cell is known, there is a very high probability that adjacent grid cells will have similar numbers of beetles; i.e., the beetle counts of neighboring sampling units are highly correlated. This is equivalent to saying that the beetle counts are not independently distributed over the study area. For reference, Figure 9B depicts one of many possible configurations of beetle abundance that are randomly (i.e., Poisson) distributed in space.

The statistical procedures discussed in this appendix require spatially referenced sampling units to be distributed like those shown in Fig. 9B. When this is so, knowledge of the count in a specific grid cell does not provide any information on the counts in neighboring grid cells. By contrast, the closer any two grid cells in Fig. 9A are to one another, the more similar their beetle counts. Thus adjacent gird cells provide redundant information regarding the distribution and abundance of beetles in the study area. A sample of 30 grid cells that are far enough apart from one another to be uncorrelated will provide more information on the variation in the population distribution than will a sample of 30 that are close together. Likewise, any estimate of population mean or median abundance from a sample of uncorrelated grid cells will be unbiased whereas estimates from the correlated grid cells will be biased towards the particular subset of grid cells that are autocorrelated.

Fig. 9. Spatial distribution of beetles (*Dychirius globosus*) on a gridded study area. Grid cells are 40 m on a side. (A) Actual observed distribution in spring 1975 (μ = 28.5, σ =58.0) (B) Randomly reassigned distribution in spring 1975 (μ = 28.5, σ =58.0).



One obvious approach to diagnosing autocorrelation in a sample would be to estimate the degree to which neighboring sampling units are correlated or alternatively independent from one another. But first some criteria are needed as to what constitutes a neighbor. In fact the degree to which any two sampling units will be spatially correlated is a function of how far apart they are. Thus it is conventional to estimate a series of autocorrelation or variance values based on the distance between neighboring sampling points. These distances are called nearest neighbor distances. Figure 10 is a plot of the variance (vertical axis) among grid cells that are 1,2,3, ….12 grid cells apart. Spatial analysts call such nearest neighbor distances, **lags** (horizontal axis). In this case, a lag unit is equivalent to the length of the side of a grid cell; i.e., 40 meters. The curve labeled "Observed" is a plot of the nearest neighbor variances of grid cells from the study area (Fig. 9A) while the nearest neighbor variances of the "Random" curve were computed from the simulated independent data shown in Fig. 9B. The horizontal reference line marks the value of the overall sample variance (3368.6).

The graph, called a **variogram**, confirms that the variability in the beetle counts from the study area tends to increase with the distance of a sampling unit from it s nearest neighbor. Conversely the closer two grid cells are to one another, the more similar are their beetle abundances, hence the smaller their neighbor-neighbor variability. By comparison, the variability among grid cells from the randomly distributed population does not vary significantly with lag distance; moreover, it never deviates much from the overall sample variance, a pattern characteristic of spatially independent data. The sharp inverted J-shaped pattern of the variogram of the observed data is emblematic of a sample with serious autocorrelation.

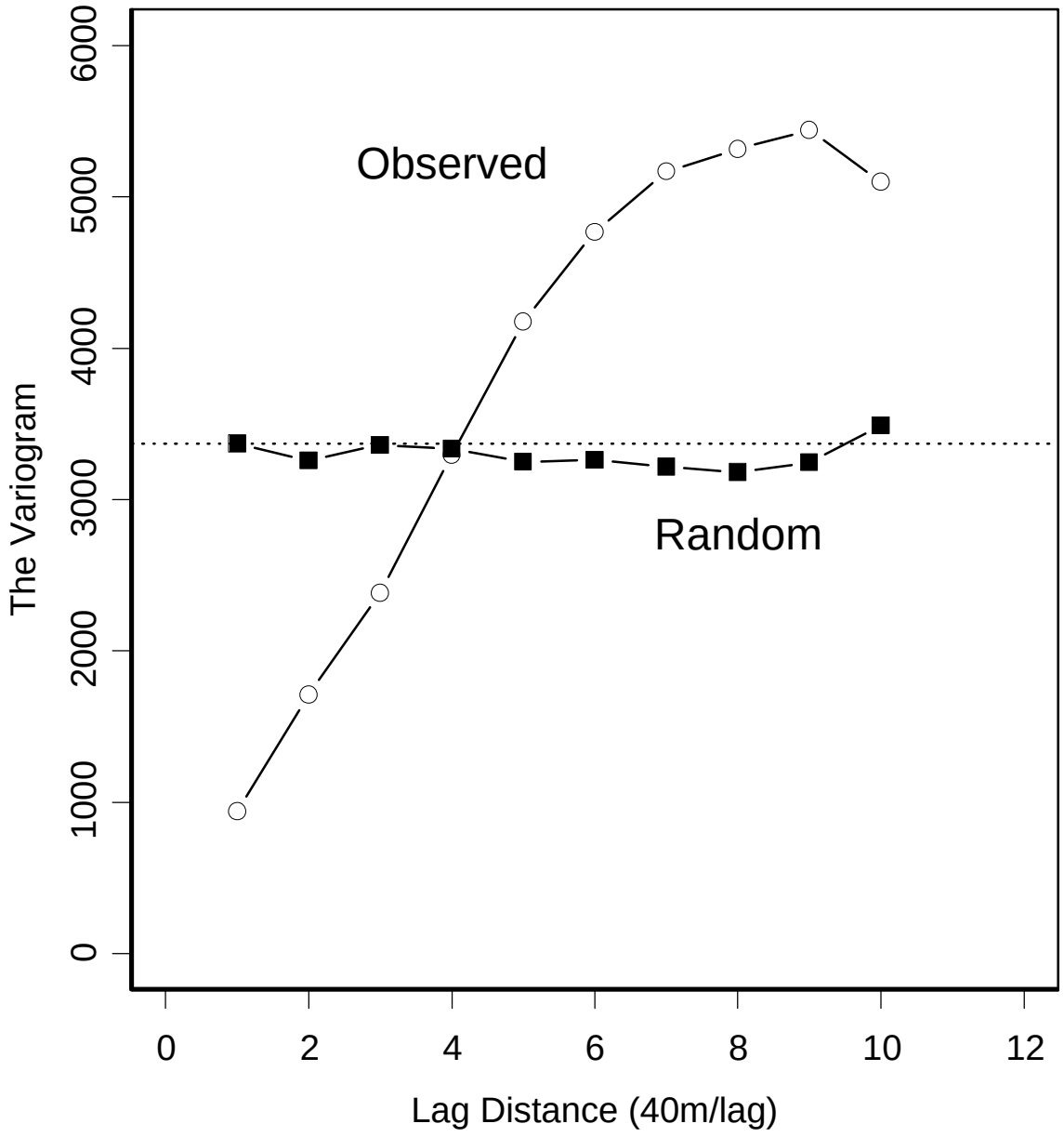
Two important conclusions can be drawn from these variograms: (1) Because the variance is an increasing function of lag-distance, the beetle count data are autocorrelated and (2) the negative effects of the autocorrelation do not extend beyond four lag-units (160m). The last conclusion is based on the lag distance associated with the intersection of the plot of the observed variogram with the overall sample variance reference line, which of course is the variance one would expect from a random independent sample.

Once spatial autocorrelation of the magnitude observed for the beetle data has been diagnosed, statistical tests and confidence intervals of the sort described in this appendix are not valid. Although there are number of methods available for the analysis of spatially and/or temporally correlated data, they are quite complex mathematically and require the assistance of an experienced spatial data analyst, using specialized software (e.g., ArcView Spatial Analyst or SPLUS SpatialStats). For water quality attainment studies, the best strategy will almost always be to construct environmental sampling designs in such a way that spatial correlation is minimized. This can be accomplished most easily be variogram analysis of appropriate pilot study data. For example, Hengeveld could have used the above results to design a fall sampling protocol wherein grid cells could not be selected if they were within 160 m of a previously selected grid cell.

C.2.5 EDA Example 4: Assessing Temporal Independence

The problems associated with temporal autocorrelation are similar to those of spatial autocorrelation, with the notable exception they occur in only one dimension (i.e., time). Just as

Fig. 10. Variograms comparing changes in the variance of beetle abundance with distance between grid cells for the observed field data (open circles) and the randomly reassigned counts (solid squares). The broken horizontal reference line marks the overall sample variance (58^2).



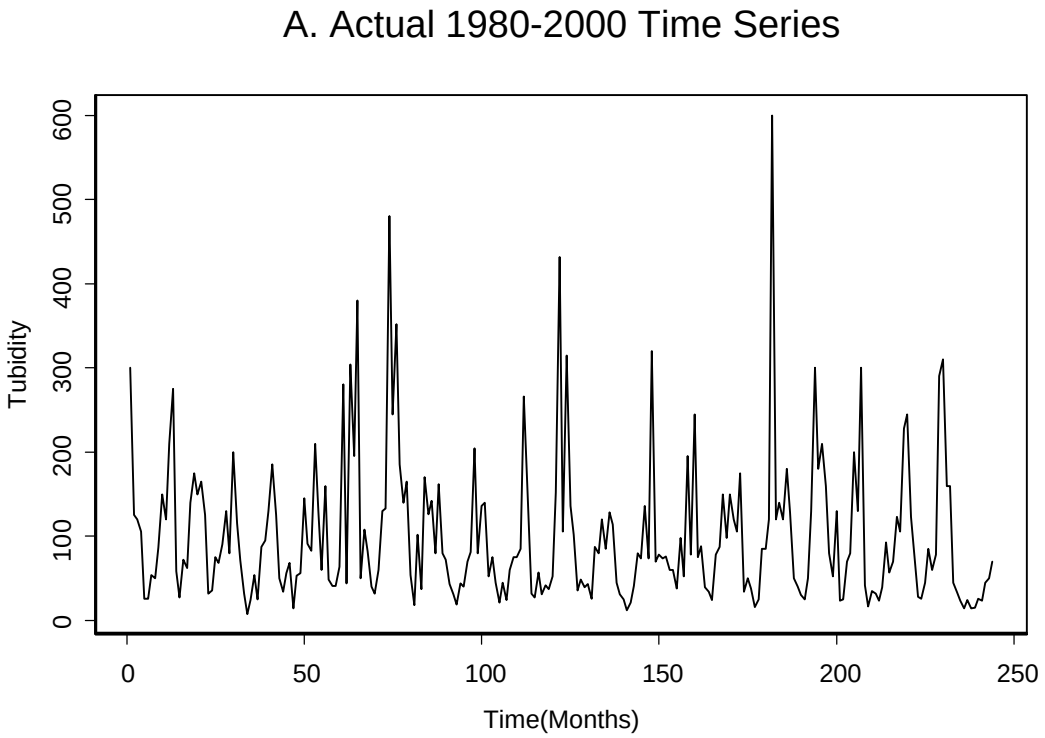
units sampled close together in space tend to be alike, so do units that are sampled close together in time. Not surprisingly then, the graphical techniques that are employed to diagnose temporal autocorrelation are similar to those used to assess spatial autocorrelation. As pointed out in Example 4, increasing correlation implies decreasing variability and *vice versa*. However, while it is customary to examine spatial autocorrelation by plotting the complementary variance relationships in variograms, time series analysts traditionally plot temporal autocorrelations directly in graphs called **correlograms**. In this example, correlogram analysis of patterns of temporal autocorrelation will be illustrated for the monthly turbidity data that were analyzed in Example 1 (Table 3).

The monthly turbidity time series is plotted in Fig. 11A. The series starts in January 1980 and terminates in April 2000. The plot reveals December-January spikes and corresponding May-July troughs that reoccur annually. This pattern suggests that seasonal factors (e.g., precipitation) may be influencing turbidity readings in the Mermentau River. Within any 1-year cycle it appears that monthly values tend to increase just before the spikes and decrease just after them. This suggests that measurements adjacent to the midwinter spike tend to be more alike than non-adjacent months. Thus the time series provides strong visual evidence of both seasonality and autocorrelation. For reference, a randomly ordered time series was generated by randomly reassigning month/year dates to the 244 monthly turbidity values, resorting and finally replotting them (Fig. 11B). The result is a chaotic series of random spikes and troughs with no apparent annual cycling. The confidence intervals and statistical tests described in the appendix assume a pattern similar to Fig. 11B and perform poorly for data with patterns like Fig. 11A.

Figures 12A and 12B are the correlograms which were computed for the actual and random time series in Figs. 11A and 11B, respectively. Each correlogram is a plot of the correlation (vertical bars) among monthly turbidity values that are 1,2,3, ….60 months apart. As in the variograms, these interval distances (horizontal axis values) are called lags. The first lag is zero; its associated autocorrelation bar is a measure of the correlation between each monthly turbidity value and itself, which of course is 1.0. The vertical bar above second lag is a measure of the correlation between each monthly turbidity value and that of the next month. Of particular interest are the correlations between turbidity values that were taken twelve months apart. These are the January 1980-January 1981, February 1980-Feberuary 1981, December 1980-December 1981 correlations, as well as the corresponding monthly correlations across any pair of successive years. Likewise, lag 24 correlations are the correlations between monthly turbidity values that are taken 2 years apart, lag 36 correlations are the correlations between monthly turbidity values that are taken 3 years apart, and so on. The two broken lines are the 95% confidence limits about the zero autocorrelation value. Thus any vertical bar that does not extend beyond the broken lines indicates an autocorrelation that is not significantly different from zero; i.e., such values denote temporal independence.

Several relationships are clarified by the two correlograms in Fig. 12. First, while there is no significant correlation beyond lag zero for the randomly resorted data (Fig. 12B), there is a consistent, recurring pattern of significant, alternately positive and negative, autocorrelations in the actual field data (Fig. 11A). The pattern in the field data repeats itself regularly at twelve-month intervals. Within a given year, the alternating positive and negative values simply reflect the up and down seasonal oscillation of the turbidity values in the Mermentau River. Significant

Fig. 11. Time series plots of 244 monthly turbidity measurements from the Mermantaut River, January 1980- April 2000. (A) The actual field data. (B) The field data randomly reordered in time.



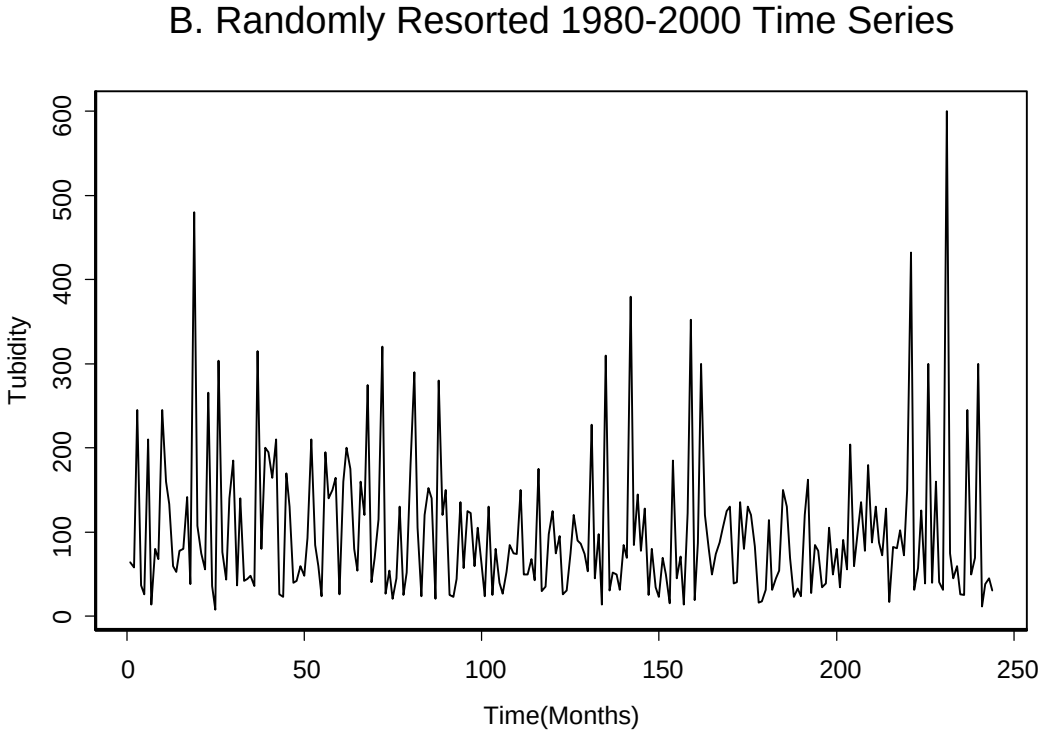
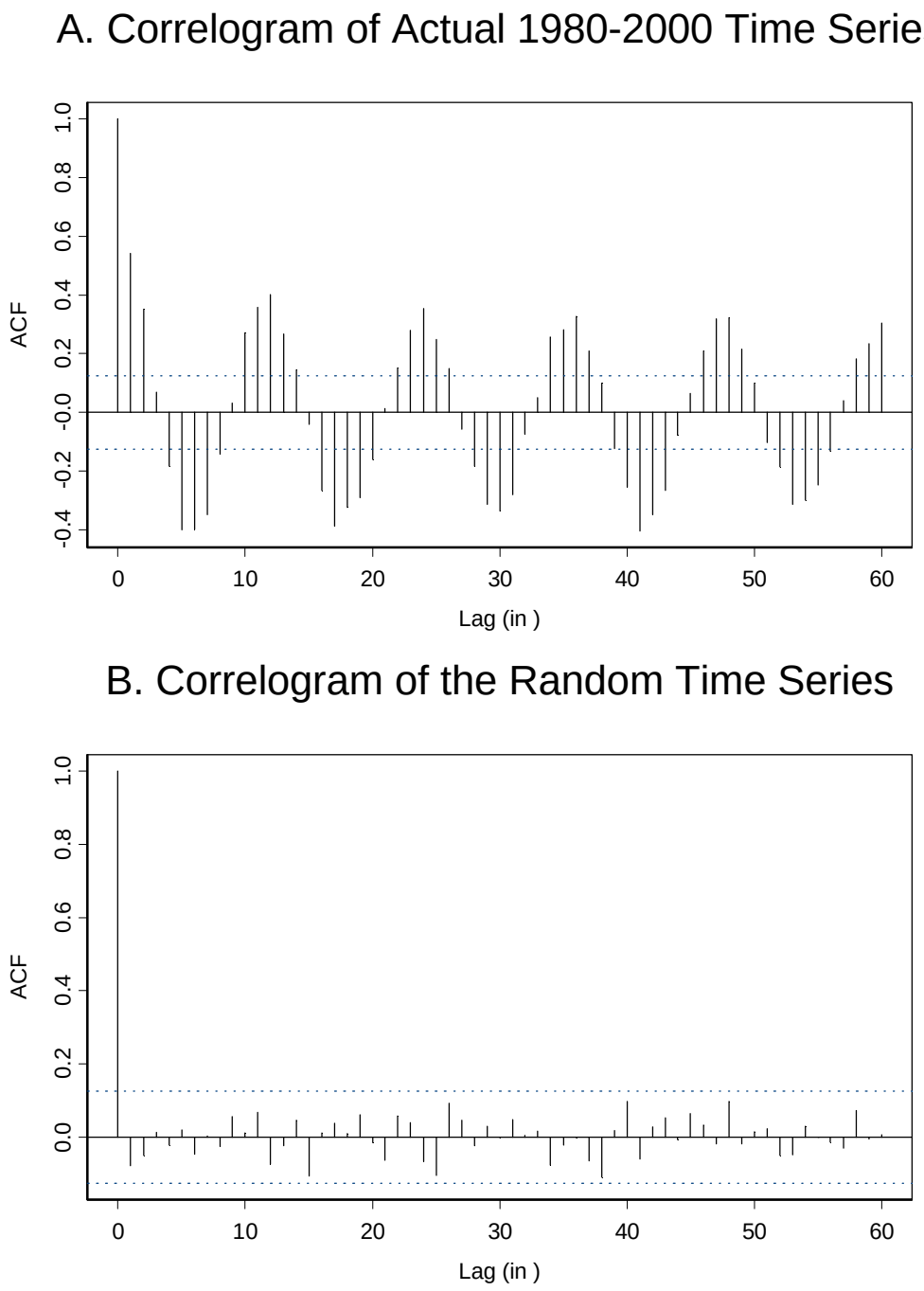


Fig. 12. Correlograms showing changes in the autocorrelation of turbidity measurements with vs. length of the time (lags) between measurements for the observed field data (A) and the randomly reordered data (B). Broken horizontal reference lines are the 95% confidence limits around zero autocorrelation; bars within them indicate temporal independence at those lags.



positive autocorrelations occur for lags (i.e., intervals) of 1, 10, 11 and 12 months with the highest correlations associated with the twelve-month lags. Significant negative autocorrelations occur for lags of 4, 5, and 6 and 7 months with the largest negative correlations associated with turbidity readings that were taken 5 or 6 months apart. Near-zero autocorrelations occur among measurements taken 3, 8, or 9 months apart.

The correlogram of the field data (Fig. 12A) suggests that analysis and/or data collection should be confined to data that have been collected at 3, 8 or 9-month intervals. This poses some serious problems to regulators who want to base decisions on three-year evaluation periods.

Since shorter sampling intervals yield larger sample sizes, the optimal choice is a 3-month interval, but this only yields twelve samples per three-year period. As will be discussed in the next section, small samples generate large decision error rates for any of the statistical tests and confidence intervals discussed in this appendix. On the other hand, autocorrelated data tend to produce biased results and hence erroneous decisions. So, what should be done? There are several alternative solutions to this common environmental data problem; one of the simpler solutions will be demonstrated in Appendix D (Section D.2).

C.3 Data Quality Assessment: Hypothesis-testing and Estimation

C.3.0 Use Of EDA and DQOs to Select an Appropriate Hypothesis-Test or Estimator

A rigorous DQO process should have included the specification of the statistical tests that are appropriate to supporting the water quality attainment decisions that motivated the study. In Step 3 of the DQA, the task of the analyst is to review, revise, and if necessary replace these tests in light of the results of the EDA. Thus this step is essentially a more informed reenactment of DQO steps 5 and 6. In Sections C.1.1-C.1.5., these two steps were discussed from the perspective of confidence interval construction. Although there is a strong natural connection between confidence intervals and statistical hypothesis tests, there are some important differences. In Sections C.3.1 – C.3.3 we will explore these differences and in Appendix D provide guidance on how to use parametric and nonparametric hypothesis tests to compare a sample mean or proportion to an environmental standard in a water quality attainment context.

C.3.1 Hypothesis-Testing Basics

Statistical hypothesis testing is broadly applicable to many situations beyond what are described here. In fact, the present application is an example of the simplest of all statistical testing scenarios: the so-called one-sample question. **One-sample tests** are typically used to evaluate hypotheses about whether some sample statistic (e.g., a mean, median or percentile) is equal to or exceeds a threshold such as a water quality standard. Tests useful for comparing two samples (e.g., a sample from an ambient or treated population to one from a reference or a control population) and tests for simultaneous comparison of samples from several populations (e.g., ANOVA and linear regression) are covered in *Guidance for Data Quality Assessment EPA QA/G-9,* in *Biological Criteria: Technical guidance for Survey Design and Statistical Evaluation of Biosurvey Data* (EPA 822-B-97-002*),* and in many standard statistical texts (e.g., Steel et al. 1996; Millard and Neerchal 2000). All of the tests described in this document can be implemented with standard commercial software (e.g., SAS and SPLUS); most of the additional tests described in *Guidance for Data Quality Assessment EPA QA/G-9,* can be carried out with SPLUS EnvironmentalStats.

Hypothesis testing is motivated by the need, during the decision-making process, to account for uncertainty in data collected by a **probability-based sampling design.** There are several possible sources of uncertainty including:

1. Sampling variation specific to the design employed to collect the data
2. Intrinsic natural (e.g., genetic or behavioral) variation among population members
3. Temporal and/or spatial variation
4. Measurement and/or laboratory error
5. Model misspecification error (e.g., in Monte Carlo risk assessments)

In a typical water quality attainment study, we usually evaluate either of two pairs of one-sided null and alternative hypotheses. The scientific objective is to compare an unknown population parameter against a specified value that prior studies have indicated is a threshold for

desirable/undesirable environmental outcomes. Depending on the situation the hypothesis pair may be either,

$$H_0 : q \leq q_0 \quad vs. \quad H_a : q > q_0 \tag{2}$$

or,

$$H_0 : q \geq q_0 \quad vs. \quad H_a : q < q_0 \tag{3}$$

where θ = the population parameter of interest
θ_0 = the fixed criterion value that the population parameter is compared against

Equation 2 evaluates an upper one-sided alternative hypothesis, while Eq. 3 evaluates the complementary lower one-sided alternative. As an example of the application of the former, we may want to evaluate the null hypothesis that the mean sediment concentration of pesticide X is less than or equal to 20 µg/Kg vs. the alternative that the sediment concentration exceeds 20 µg/Kg.

Having selected the population parameters of interest (e.g., the mean concentration of a pollutant) and chosen null and alternative hypotheses appropriate to the water quality attainment decision under consideration, the next step is to choose a statistical test which can be used to determine which hypothesis (i.e., H_0 or H_a) is better supported by the sample data. Statistical tests are simple mathematical models that predict the distributions of test statistics when the null hypothesis is true. These are the distributions one would expect to obtain from conducting thousands of surveys or experiments and plotting the frequencies of the computed test-statistics. These distributions, called **sampling distributions**, reflect the uncertainty in the sample estimates of the values of the test statistic.

Statistical tests are commonly named for their sampling distributions (e.g., t-test, chi-squared test). The test statistics themselves are usually simple algebraic functions of the sample statistics. For example, the test statistic for the test against either one-sided alternative for comparing the mean of a continuous random variable, such as the sediment concentration of pesticide X, to a fixed value such as a 20 µg/Kg environmental standard is a function of the sample size (n), mean ($\bar{x}$) and variance (s_x^2):

$$\frac{\bar{x} - 20}{\sqrt{s_x^2 / n}} \tag{4}$$

Statistical theory insures that when certain assumptions hold and the null hypothesis is true, the sampling distribution of the test-statistic shown in Eq. 4 will be a t-distribution with n-1 degrees of freedom (df). Thus the associated statistical test is called a t-test. There is actually an entire family of t-distributions, each with a different df.

The sampling distributions of the test statistics should not be confused with the population distributions from which the samples have been collected. Whereas the latter (e.g., Fig. 5a) are the natural distributions of the animals or water quality factors under study, the former are statistical models of the behavior of statistics calculated from samples of the natural populations. Some statistical tests, called **parametric tests** (e.g. t-tests) require that the natural populations be

normally distributed, while others, called **nonparametric tests** (e.g., chi-square tests), make no assumptions about the distribution of the natural populations.

All statistical hypothesis tests are mathematical models. In the case of t-tests, chi-square tests and F-tests (i.e., tests associated with ANOVA and regression models), the distribution of the test statistic computed from the sample data is modeled as (respectively) a t-distribution, a chi-square distribution or an F-distribution. Like all models, the validity of the predicted distributions depends on assumptions made about the underlying processes that are being modeled. In the case of one-sample t-tests for the population mean, the following assumptions are made:

1. The variable being analyzed is a continuous random variable that is normally distributed in its target population.
2. The sampling units used to compute the sample mean from which the t-statistic (Eq. 4) was computed were independently distributed in the target area (i.e., there is no temporal or spatial autocorrelation among the sampling units).
3. There was no systematic error associated with measuring the response on the sampling units (e.g., pH meters and/or laboratory assays were correctly calibrated and applied).
4. The null hypothesis is true.

The distribution of the test statistic under the null hypothesis is the basis for determining whether the data support the null hypothesis or the alternative. For example if we have a sample of 30 sediment sampling units, the expected distribution of t-statistics under the null hypothesis is a t-distribution with df=29. Ninety-five percent of the t-statistic values in such a distribution are less than 1.699. Thus if the t-statistic value computed from our sample has a value of (say) 8, then we conclude that there is less than a 5% probability that our sample came from such a t-distribution. This suggests that one or more of the above four following assumptions is *not* true. If we have previously verified the first three assumptions (e.g., by application of the appropriate DQOs and EDA methods in Sections C.2.1- C.2.5), we can conclude that our sample *does not* support the null hypothesis. If we have not verified the assumptions, we cannot draw any conclusions from the t-test. Rejection of the null hypothesis provides evidence in favor of the alternative that the sediment concentration of pesticide X exceeds the environmental standard of 20 μg/Kg.

The probability used as the cutoff for accepting or rejecting the null hypothesis is called the **significance level**. By declaring a significance level of 5%, we are saying that even though there is a 5% probability that a t-statistic ≥ 1.699 could have come from the null t-distribution, this probability is so small that we don't think it is reasonable to believe that the data actually came from a target area whose pesticide X concentration was ≤ 20 μg/Kg. On the other hand, there *is* a 5% chance that it could have. Thus the significance level is just the Type I error rate (α) that the investigator decided in DQO step 6 he would be willing to live with.

C.3.2 Types I and II Error Rates and Statistical Power

In this section graphs of the distribution of the test statistic when the null hypothesis is true will be compared to its expected distribution when it is false. The graphs will be used to clarify the relationships between the two types of decision error and the statistical power of the hypothesis

test. The important effects of three attributes of the sample on the power of the t-test also will be explored and illustrated through the use of graphs. These factors are: the sample size, the sample variance, and the ± difference between the sample estimate of the population parameter (e.g., the pollutant mean concentration) and the criterion value that is the basis for an attainment/impaired classification. As will be seen, the observed power of any statistical test (*not* just the t-test) is the result of a complex interaction of these three factors and the investigator's choice of the null and alternative hypotheses.

We will begin by examining the hypothesized distributions of pesticide X in the sediments of the target area and the distributions of the corresponding t-test statistic under both the null and alternative hypotheses. Recall that alternative hypothesis is open ended; i.e., it states only that the true sediment mean concentration is greater than the criterion value of 20 μg/Kg. Thus any value > 20 will be consistent with H_a. Figure 13 shows the hypothesized distributions of the sampling unit concentrations of pesticide X as they are expected occur in the sediments of the target area if H_0 is true (normal: $\mu = 20$ and $\sigma = 1$) and a possible alternative distribution if H_a is true (normal: $\mu = 21$ and $\sigma = 1$). Using the sample mean and standard deviation, a t-statistic for testing the null vs. the alternative hypothesis can be calculated with Eq. 4.

The influence of sample size on the expected distributions of this t-statistic under the null and alternative hypotheses are shown in Figs. 14a and 14b . The distribution of a t-statistic under the alternative hypothesis is called the **noncentral t-distribution**. Based on a sample size of 30 sampling units (Fig. 14b), the two distributions are quite distinct with only a sliver of overlap. In contrast, the distributions of test statistics computed from the much smaller sample size of 10 (Fig. 14a), overlap considerably. By analogy, we can think of the distributions as distant mountains viewed through a telescope. While the two peaks are distinct when observed through a powerful telescope (Fig. 14b), they appear to merge when seen through a weaker telescope (Fig. 14a). The heavy vertical line in Fig. 14a marks the point at which the t-statistic values of the alternative distribution are less than the 95^{th} percentile of the null t-distribution. The noncentral t-values that lie to the left of this line represent t-test outcomes that are erroneously regarded as evidence that the data come from the null distribution. This proportion of the alternative distribution is the Type II error probability (β); it is the part of the "alternative mountain" that can't be distinguished from the "null mountain." The complementary proportion (1-β) of the alternative distribution to the right of the 95^{th} percentile of the null, is the part of the alternative distribution that is *correctly* distinguished from the null distribution. The larger this proportion, the greater the resolution of the test. Thus 1-β is a measure of the **power** of the test to correctly identify t-statistic values that support the alternative hypothesis. It is clear from Fig. 14a why, for fixed sample size and variance, decreasing the α value increases the β value; decreasing α moves the heavy vertical line to the right, causing more of the alternative to "merge" with the null distribution.

The influence of the sample standard deviation (σ) on the power of the t-test is illustrated in Figures 15a and 15B. All the distributions shown in Fig. 15 are based on sample sizes of 30 (df=29) from populations with the same standard deviation of 15. Like the distributions in Fig. 14b, the null and the noncentral t-distributions in Fig. 15a have means 20 and 21, respectively. Thus the only difference between the Fig. 14b and 15a distributions is that the former have $\sigma=1$ while that latter have $\sigma=15$. However, this 15-fold increase in variance results in a decline

Fig. 13. Distributions of two populations of sediment sampling units, both of which are normal with standard deviations of one, but with different means (solid line = mean of 20; broken line = mean of 21). The null hypothesis states that the former is the true population distribution, while the latter is one of several possible population distributions that could occur if the alternative hypothesis is true.

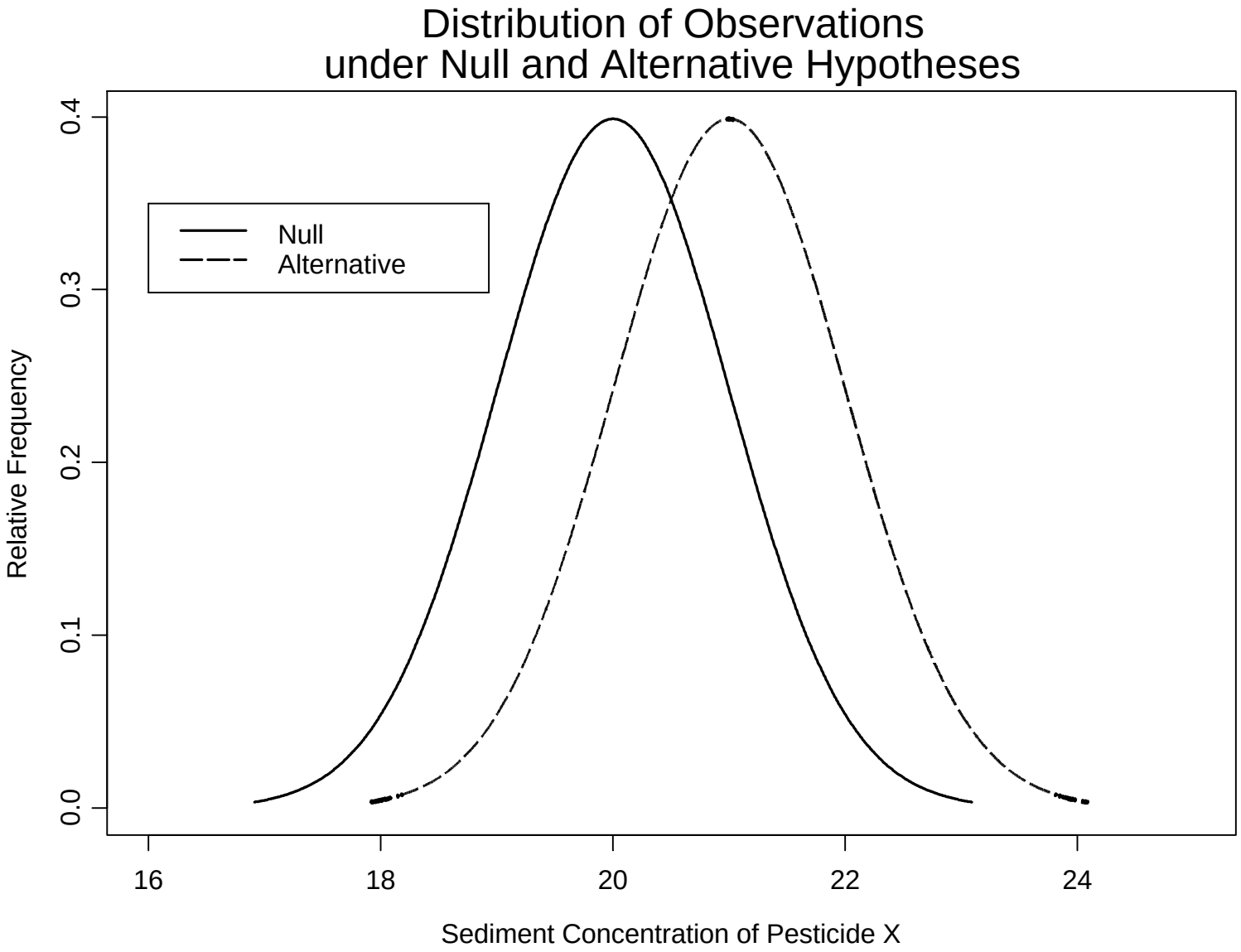
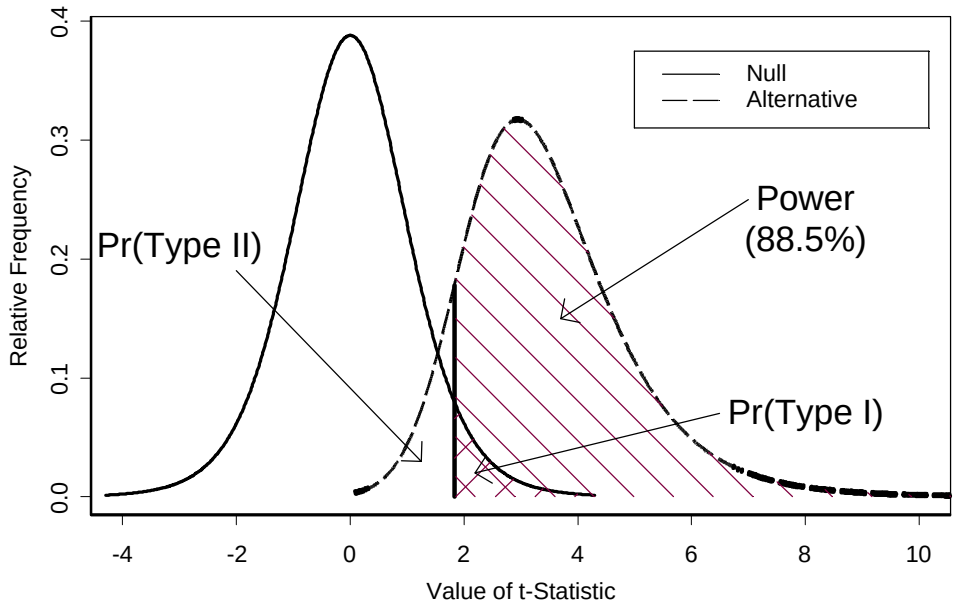


Fig. 14. Sampling distributions of t-statistics with both their population standard deviations and the effect size equal to 1 when the null hypothesis is true (solid line) vs. when the alternative hypothesis is true (broken line). (A) compares the two distributions when the sample size is 10; (B) compares them when the sample size is 30.

A. Null and Alternative t-Distributions (n=10, σ =1, δ =1)



B. Null and Alternative t-Distributions (n=30, σ =1, δ =1)

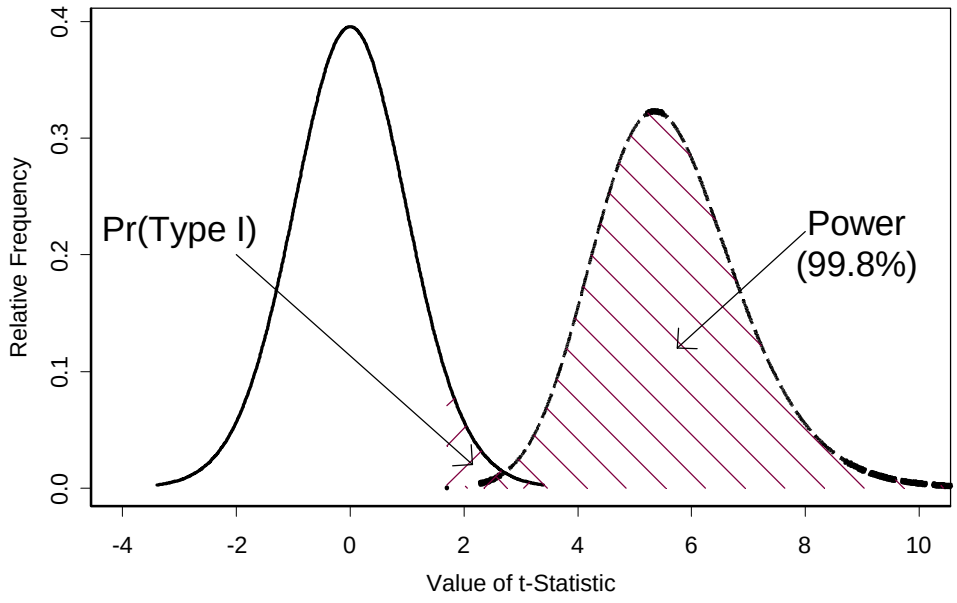
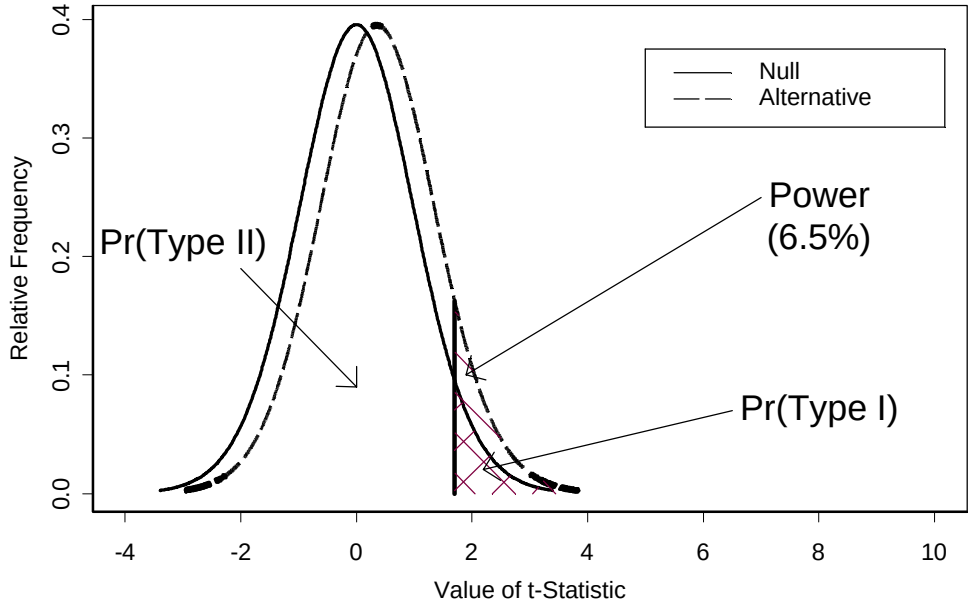
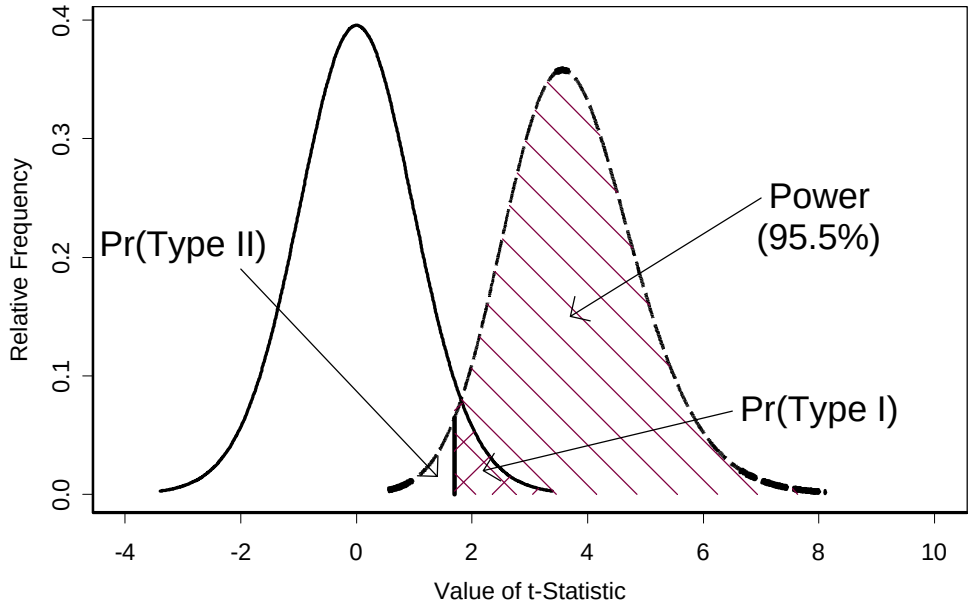


Fig. 15. Sampling distributions of t-statistics with population standard deviations of 15, and sample sizes of 30 when the null hypothesis is true (solid line) vs. when the alternative hypothesis is true (broken line). (A) compares the two distributions when the effect size is 1; (B) compares them when the effect size is 10.

A. Null and Alternative t-Distributions (n=30, σ=15, δ =1)



B. Null and Alternative t-Distributions (n=30, σ=15, δ =10)



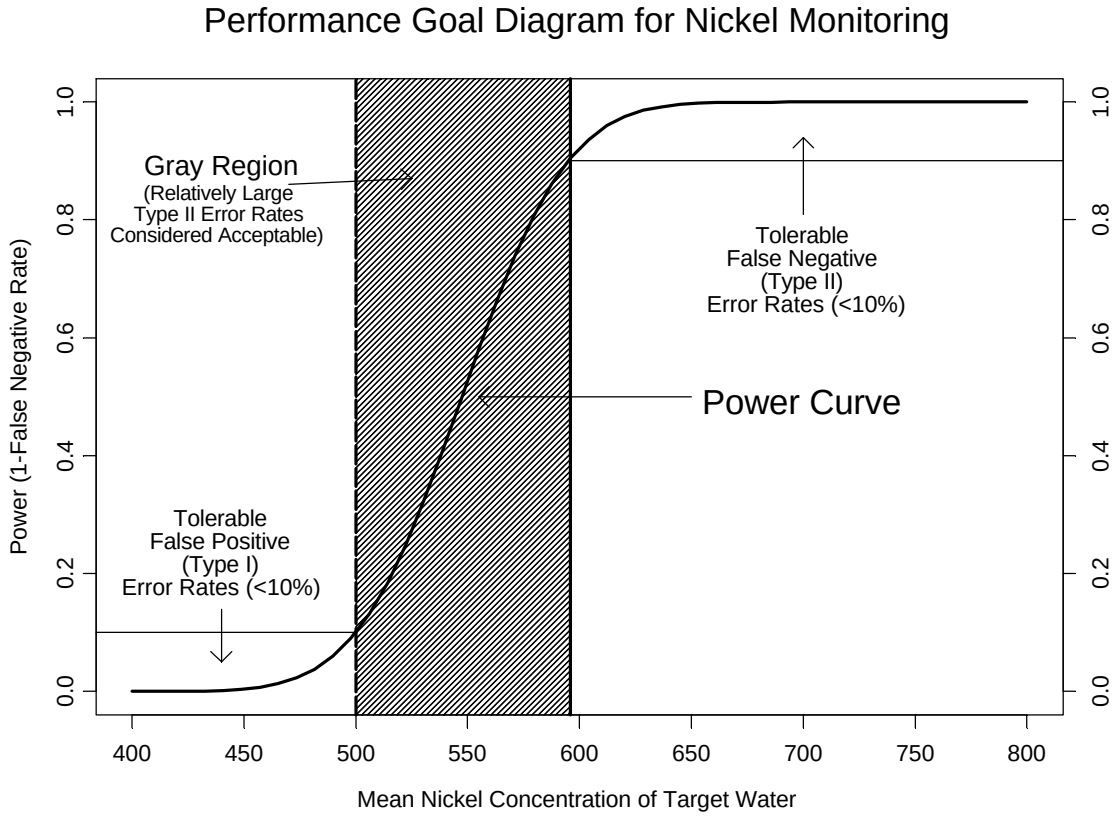
in power from 99.8% to only 6.5% for the fixed sample size of 30. As illustrated in Fig. 15a, the null and the alternative distributions occupy nearly identical locations with only a small proportion (6.5%) of the alternative distribution to the right of the 95th percentile of the null distribution. This illustrates the general principle that an increase in the variance of a sample will cause a decrease in the power of an statistical tests based on that sample.

The effect of the size of the difference between the sample mean and the criterion value on statistical power can be seen by comparing Fig. 15a with Fig. 15b1. In Fig. 15b, the noncentral t-distribution has a mean that is 10 units larger (i.e., 30) than the null distribution. This has the effect of shifting its location 10 units to the right of the null, simultaneously reducing the Type II error (i.e., the overlap with the null) and increasing the power to 95.5%. Returning to our telescope analogy, we note that if one object is located directly in front of the other, it is nearly impossible to distinguish them. But if one of the objects is shifted away from the other, it becomes discernible.

In the language of experimental design, the observed difference between a mean (or median) and the criterion value to which it is being compared is called the **effect size**. Figures 15a and 15b illustrate the general case: the smaller the effect size (usually symbolized by δ), the more difficult it is to for a statistical test to distinguish the sample mean (or median) from the criterion value; i.e., the power is low. For fixed δ and population σ, the only way to increase the power to a point where the difference becomes "visible" (i.e., H_0 is rejected), is to increase the sample size. Thus whenever a statistical test will be used to compare a sample mean or median to a criterion value, it is important decide *a priori* how large the difference (i.e., δ) must be to be "ecologically significant". Once this is done, an appropriate minimum sample size can be estimated. Statistical significance for a study designed in this manner will be a reasonable indicator of ecological significance. If there is not a consensus on how large the effect size should be for it to be ecologically or toxicologically significant, it will be difficult (perhaps impossible) to use the results of the statistical test to support a WQS decision.

Figure 16, called a Decision Performance Goal Diagram, presents a graphic summary of the relationship of effect size to the false negative rate (i.e., Type II error rate) and its complement the statistical power associated with either a lower one-sided 90% confidence interval or a t-test with an upper one-sided alternative. In this example, the sample mean concentration of nickel is compared to a CMC 500 μg/L. The null hypothesis is that the concentration is $\leq$ 500 μg/L vs. the alternative that it is > 500 μg/L. The sample size is 30 one-liter sample units drawn from water body of interest and the standard deviation is 200. The vertical axis values are the power (i.e., 1-β) and the horizontal axis displays the reasonable expected range of the nickel concentrations in the sample. The power curve traces the power against the one-sided alternative hypothesis. Note that when the sample means are low (<450 μg/L) the false positive (i.e., Type I) error rates and the power are near zero. This is at is should be; sample means in this range do not provide much basis for rejecting H_0. However as the sample means increase from 500 to 596, the statistical power increases rapidly. This "gray" area represents the range of values where inference is unreliable; it is bounded on the left by the criterion value and on the right by the criterion value + the effect size (i.e., δ=596-500 μg/L). The right bound of the gray area is the value at which the sample size of 30 provides sufficient power to insure rejection of H_0 in $\geq$ 90% of samples. Thus Fig. 16 suggests that given a sample size of 30 sampling units and an

Fig. 16. Decision performance goal diagram for a one-sample t-test of H_0: population mean ≤ 500 vs. H_a: population mean >500, when the standard deviation =250, the sample size=30 and the minimum detectable effect size=100.



effect size of 96 μg/L, the t-test will have both good power and low false positive error rates. Once the mean exceeds 596 μg/L, H_0 will be rejected with probability $s \geq 0.90$ (i.e., false negative error rate < 0.10). From a confidence interval perspective, specifying a narrow half-width on the interval estimate is analogous to decreasing the minimum detectable effect size δ; either of these actions will decrease the width of the gray area by moving the right bound closer to the criterion value, thereby increasing the area with tolerable Type II error rates to the right of the gray zone. See Chapter 6 in "Guidance for the Data Quality Objectives Process" (EPA QA/G-4) for additional discussion of Decision Performance Goal Diagrams.

For a one-sample test, the effect size is essentially a detection limit. Although one can set a criterion such as 500 μg/L for a concentration, if one employs a one-sample t-test or one of its nonparametric counterparts, the result of the test will be depend on the whether the difference between the sample mean or median and 500 μg/L criterion is large enough to be statistically significant. The effect size tells us how much the sample mean must exceed the criterion value before the exceedance is detected with a probability of 1-α. Thus the effect size effectively resets the criterion value to a new value equal to the sum of the original value + the effect size (e.g., 500 + 96 μg/L). This adjustment is an acknowledgment of the effects of the uncertainty in the sample estimates on the Type II error rate. Perhaps, just as important, in setting of the minimum effect size, the investigator must recognize the cost constraints on the sample size, " In essence, the gray region [which is determined by δ] is one component of the quantitative decision performance criteria that is specifically used to limit impractical and infeasible numbers of sampling units" (GS-4, page 6-4).

In summary, the observed power of a statistical test is the result of a complex interaction of the sample size, the variance of the attribute being measured, the effect size, and the specified α-level. When each of the other three factors is fixed:

1. Decreasing the variance increases the power of a statistical test
2. Increasing the α-level (e.g., 0.05? 0.10) increases the power of a statistical test
3. Increasing the effect size (gray area in Fig. 16), increases the power of the test
4. Increasing the sample size increases the power of a statistical test

In general, it will not be advisable to base a WQA decision on the p-value from a hypothesis test unless the interactive effects of sample size, sample variance, effect size, and the specified tolerable Type I error rate (i.e., the α-level) on the power of the test have been carefully considered and are consistent with the DQOs and the objectives of the WQS.

C.3.3. Reversing the Null and the One-sided Alternative Hypotheses

Up to this point, our discussion of one-sided tests for comparing a population parameter θ (e.g., percent exceedance or mean pollutant concentration) to a regulatory standard k has focused on paired hypotheses of the form:

$$H_0 : q \leq k \quad vs. \quad H_a : q > k \tag{5}$$

These hypotheses lead us to define the Type I error rate as the probability of incorrectly classifying as "impaired", a body of water that attains the water quality standard. Similarly, the Type II error rate is defined as the probability of incorrectly classifying an impaired water as one that meets the standard. Because the Type I error rate (α) is fixed by the investigator or regulatory agency, its value is assured regardless of the sampling design implemented or the variability of the monitored water body. However, as shown in the preceding section and in Section D.5, the Type II error rate (β) depends on the complex interaction of the specified α, the sampling design, the sample size, the variance of the target water and on the resulting minimum detectable effect size (δ). This last point is especially vexing since δ effectively increases the minimally acceptable pollutant level from k to k+δ.

From the perspective of environmental protection, it can be argued that it is much more desirable to fix β rather than α. For example, by fixing β at 0.05, one can assure that there is a 95% probability that any water that exceeds the criterion value k will be correctly identified as impaired, regardless of the amount by which it exceeds k. However, as long as attainment of the water quality standard is the baseline condition, investigators and/or regulatory entities will not be able to control the Type II error rates.

A solution is available if the null and the alternative hypotheses are "flipped" like so:

$$H_0 : q \geq k \quad \textit{vs.} \quad H_a : q < k \tag{6}$$

The flipping of the null and the alternative results in a concomitant exchange of the Types I and II errors. Thus, under the new set of hypotheses, α becomes the probability of incorrectly classifying, as attaining the standard, a water that is in fact impaired, while β becomes the probability that a water body that meets the attainment criterion is incorrectly classified as impaired. As in the former case (i.e., Eq. 5), the regulatory agency can fix α at 0.05 or any other rate of its choosing, thereby insuring that all but 5% of impaired waters are correctly classified, regulated and/or remediated.

There are three additional benefits to flipping the hypotheses:

1. Power now becomes a compelling concern of the monitoring entity (e.g., the polluter). This will provide a powerful incentive for it to devise the best possible sampling designs employing the largest affordable sample sizes, in an attempt to prevent the water(s) from being incorrectly listed. Consequently, the general quality of WQA studies would likely be improved greatly. The currently implemented testing scenario (i.e., Eq. 5) actually supplies a disincentive for doing this.

2. The new approach (Eq. 6) avoids the inevitable problem of specifying an effect size (δ) that increases the desired criterion value (k) to a new, less protective, threshold (i.e., k+δ; see also Fig. 16).

3. The new approach provides a financial reward to a monitoring entity that has kept pollution far below the criterion, since a much smaller n is required to protect against a Type II error

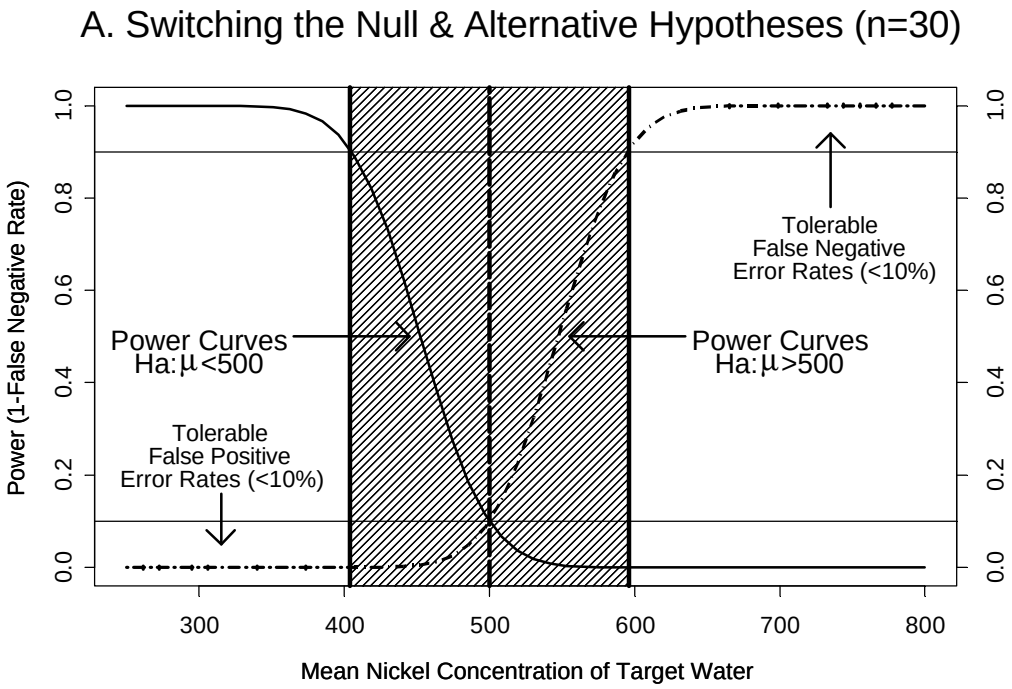
(i.e., to have high power), the farther the sample mean is below the criterion. Conversely, monitoring entities that have reason to believe the mean of the water under study is close to the standard will need to employ a much larger sample size to ensure a comparable level of power.

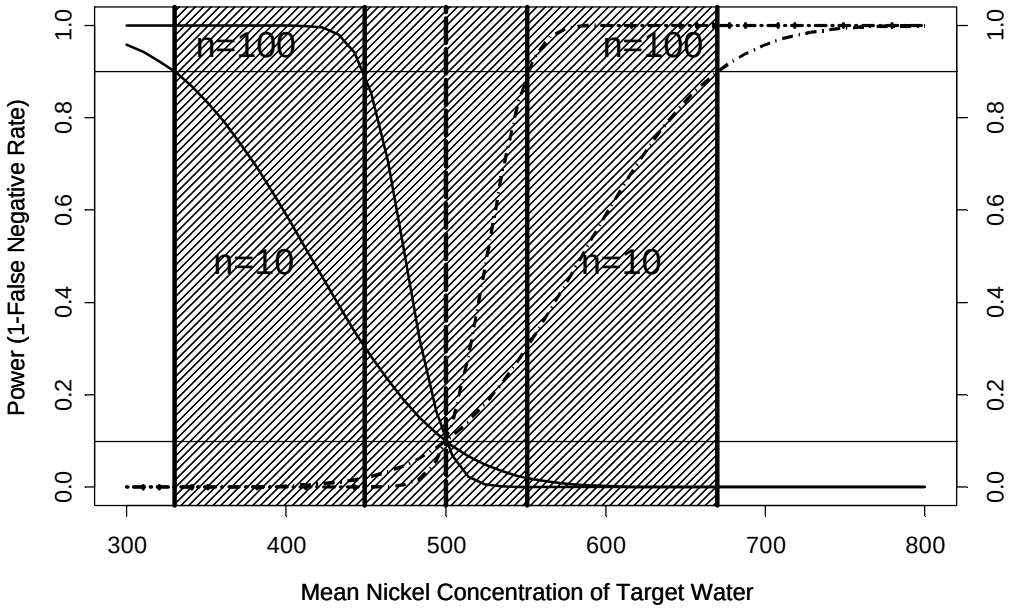
These relationships are illustrated in Figs 17a and 17b, which refer to the nickel monitoring problem described in Section C.3.2 and in Fig. 16. Figure 17a illustrates two power curves and their associated "gray regions", based on a sample size of 30 aliquots and a standard deviation of 200 μg/L. The broken line is identical to the power curve in Fig. 16; it corresponds to the null hypothesis that the mean nickel concentration in the target water is ≤ 500 μg/L vs. the alternative that the mean concentration is greater than 500 μg/l. The solid line is the power curve associated with the "flipped" hypotheses wherein the null becomes H_0: μ≥ 500 vs. the alternative, H_a μ<500. A pair of horizontal reference lines, one near the top and one near the bottom of the graph, mark respectively the zones within which the false negative error rates (i.e., Type II) and false positive error rates (i.e., Type I) are < 10%. Because the population standard deviation is 200 and we have fixed the Type I error rate at 10% and the sample size at 30 aliquots, the minimum detectable difference (δ) between the sample mean and the criterion value (k=500) is 96 μg/L. This is true for both pairs of hypotheses; thus the gray region associated with H_a μ>500 extends from 500 μg/L to 596 μg/L while the gray region associated with H_a μ<500 extends 96 units in the opposite direction (i.e., 404 μg/L - 500 μg/L). The two gray regions appear respectively to the right and left of the vertical dashed line (at k=500) in Fig. 17 a.

The interests of the regulating agency focus on controlling errors to the right of the dashed 500 μg/L vertical reference line. As pointed out above, the regulating agency can specify and control the false positive error rate. When the alternative hypothesis is H_a μ<500, the area controlled (i.e., the area below the lower horizontal reference line) will be on the right side of 500 μg/L, whereas the area that is subject to the effects of the sample size, the variability, and the minimum detectable difference (i.e., the area above the upper horizontal reference line) will be on the left side. When the specified alternative is H_a μ<500, Fig. 17 a shows that whenever the sample mean concentration of nickel exceeds the 500μg/L criterion value, the probability that the associated water will be *correctly* classified as impaired *will always be at least 90%*, (i.e., the solid curve is always below the 10% false positive reference line for nickel concentrations > 500μg/L). The left-side gray region (i.e., 404-500 μg/L) corresponds to the range of sample means that have a fairly high probability of being incorrectly classified as exceedant (i.e., the false negative rate associate with H_a μ<500), even though the observed sample mean is less than the 500 μg/L criterion. It is to the advantage of the monitoring entity that the width this region be shrunk towards the criterion value. As explained in Section C.3.2, this is usually accomplished through optimizing the sampling design and increasing the sample size.

Figure 17b further illustrates the advantage to the monitoring entity of increasing the sample size when the mean concentration is believed to close to the criterion value; i.e., for situations in which there is a reasonable probability of exceedance. The solid lines are the power curves associated with H_a μ<500 when the sample sizes (n) are 10 or 100 aliquots. Similarly, the broken lines are the power curves for the complementary alternative, H_a μ>500 for n=10 or n=100. The inner pair of vertical reference lines embrace the gray regions associated with

Fig. 17. Decision performance goal diagrams contrasting power curves for studies employing one-sample t-tests with H_a μ>500 (solid sigmoid curves) vs. those with H_a μ<500 (dashed sigmoid curves). Panel A is based on a sample size (n) of 30; Panel B compares power curves based on n=10 to those based on n=100. See discussion, pages 25-26.





n=100, while the outer pair bound the gray regions associated with n=10. Note that the solid vertical reference lines intercept the upper horizontal reference line at the points where each power curve crosses the 10% false negative error rate boundary and that regardless of which sample size is chosen, the false positive rates associated with H_a $\mu<500$ will always be less than 10% (lower horizontal reference line). A monitoring entity that is confident that the mean concentrations of nickel in the target water are lower than 300 μg/L will be satisfied with a sample size of 10 aliquots. However, if it suspects that mean levels are more likely in the vicinity of 450 μg/L, it will probably want to invest in a sample size of at least 100 because of the relatively high probability of erroneously listing the water if a smaller sample (e.g., 10 aliquots) is used. The choice of n=10 or n=100 will not be a matter of concern to the regulating agency since, in either case, the environment will be protected against falsely declaring attainment of the criterion.

Although it might appear that the balanced α-β approach described in Section D.5 would not be effected by reversing the null and alternative hypotheses, it fact there would likely be beneficial consequences because of the interest, on the part of the monitoring entities, in keeping β as low as possible. This would make it easier for the regulatory agency to obtain broad support for lowering the values of *both* α and β, thus improving the overall quality of WQ assessments.

As is the case for the currently implemented approach, one can base a WQA decision associated with the reversed hypotheses (Eq. 6) on 1-sided 100×(1-α)% confidence intervals. There are two possible 1-sided confidence intervals for the population mean concentration μ of a pollutant, the lower 1-sided 100×(1-α)% confidence interval,

$$\left[\bar{Y} - t_{n-1,1-a} \times \frac{s_y}{\sqrt{n}}, \infty\right] \qquad (7)$$

And the upper 1-sided 100×(1-α)% confidence interval,

$$\left[-\infty, \bar{Y} + t_{n-1,1-a} \times \frac{s_y}{\sqrt{n}}\right] \qquad (8)$$

In the case of the alternative hypothesis in Eq. 5, a body of water will be listed only if k is *not* contained within the lower 1-sided 100×(1-α)% confidence interval. This can only occur if the lower bound is greater than k. Under these hypotheses, a mean that is considerably greater than k may still be judged to come from attaining waters if the lower bound of the confidence interval is less than k.

The upper 1-sided 100×(1-α)% confidence interval (Eq. 8) is appropriate for assessing attainment/impaired status whenever the hypothesis pairs in Eq. 6 are to be tested. In this scenario, only waters whose mean has an upper bound that is less than k will be classified as attaining the standard. Since, by definition all sample means that are greater than k must have upper 1-sided confidence limits > k, all bodies of water with sample means >k will be listed with at least 100×(1-α)% confidence. Thus regardless of whether one chooses to base their WQA decision on confidence intervals or on hypothesis tests, the null and alternative hypotheses of Eq.

6 will always lead to a decision wherein impaired waters will be listed with at least $100\times(1-\alpha)\%$ confidence.

C.3.4. Data Quality Assessment Case History: Monitoring dissolved oxygen (DO) downstream from an agricultural operation

This example concerns the comparison of the proportion of samples that fail to meet a water quality criterion, to the maximum proportion allowed by state or federal water quality standards. The example is presented in a continuous format to show the five-step DQA process and its relationship to the DQO process that preceded it.

0. Background

In January 2000, a 3-year dissolved oxygen monitoring program was initiated at a sampling station located 0.60 miles downstream from a large commercial hog operation that had a history of manure lagoon overflows. This example demonstrates the application of the DQA process to make an interim decision based on a statistical analysis of data from the first full year of monitoring. In this case, the study was formulated, planned, and implemented through a rigorous DQO process that is illustrated in the DQO example on pages 15-22. Two additional DQA case histories will demonstrate how to apply DQA principles to the analysis of data from studies that were not designed through the DQO process.

1. Review the Study DQOs and Associated Sampling Design

Review the study Objectives. The basic research problem has been described in the preceding background discussion. The data for the first year included 34 DO measurements taken at 11-day intervals beginning on January 2, 2000. All of the scheduled evaluations were made; there were no missing data (Table 5). The unit of analysis is the area-adjusted mean DO measured at the study site on each of the 34 evaluations.

Translate the Objectives into Statistical Hypotheses: The baseline condition assumed that no more than 10% of the 34 samples had area-adjusted mean DO values < 5.0 mg/l. The alternative condition was that more than 10% of the samples failed to attain this DO criterion. Therefore the null and the alternative hypotheses are:

H_0: the proportion of the 34 samples with mean DO < 5.0 mg/l is ≤ 0.10
H_a: the proportion of the 34 samples with mean DO < 5.0 mg/l is > 0.10

The DQO process specified a gray region that was bounded on the left by the action level (0.10) and on the right by 0.25 (Fig. 4). A decision was made to accept decision errors within the gray region. For example, if one erroneously concluded that a set of samples with 15% of its area-adjusted mean DOs < 5.0 mg/l had attained the standard (10%), the resulting error would not have serious consequences because the proportion of attaining samples would still be quite high (85%). Outside the gray region, the acceptable false negative and false positive error rates were constrained to be ≤ 0.15 (Fig. 4). With a sample size of 34 and these specifications, the occurrence of 6 or more DOs < 5.0 mg/l (i.e., $\geq$17.6%) would result in rejection of the null

hypothesis with a false positive rate of ≤ 0.12. Conversely, if 5 or fewer non-attainment means were observed, the null hypothesis would be accepted with a false negative rate of ≤ 0.15.

2. Conduct a preliminary Review of the Data

The 34 area-adjusted mean DO values are tabulated, by sampling date, in Table 5. DO values that are below the action level are marked with an asterisk. Twenty out of 34 means (59%) were below the action level. Except for a brief period in January – May, it appears that DO values at the study site regularly and frequently fell below the 5.0 mg/l action level.

A frequency distribution of the DO values is shown in Fig. 18a. The graph was made by subdividing the sample data into 14 groups based on the their DO values. Each bar is centered on the midpoint of the range of DO values in each group; the height of the bar denotes the number of samples in each subgroup. From inspection of Fig. 18a, it is obvious that the majority of the distribution of DO values is less than the 5.0 mg/l action level. There is little doubt that the10% non-attainment criterion was exceeded at the monitoring site during 2000. Nonetheless, it is desirable to incorporate some estimate of uncertainty (due to sampling error) into the decision-making process. A statistical test of hypotheses will be employed for this purpose.

3. Select the Statistical Test

Selecting the Test. Although the original data were means, the criterion is written in terms of the proportion of those means that do not attain the minimum DO concentration of the action level. The proportion itself is just the number of non-attainment means (20) divided by the total number of samples (34). This proportion (0.59) is compared against the criterion value (0.10). However, the comparison should take account of the uncertainty that arises from sampling the water at the station during 2000. In order to do this, a mathematical model of the random sampling variation in the proportion of non-attainments must be used. Because the proportion measures the occurrence of one of two possible outcomes for each sample (i.e., attainment or exceedance of the 0.10 criterion), the discrete binomial distribution is the appropriate model.

There are two statistical tests available to test the null hypothesis (attainment) vs. the alternative (exceedance): the exact binomial (binomial test) and the normal approximation (Z-test). Because the Z-test is based on an approximation to a discrete distribution by a continuous distribution, a continuity correction is sometimes made to the Z-test statistic before computing the p-value. The normal approximation gives very close agreement to the exact binomial for sample sizes greater than 50 and very poor agreement for sample sizes less than or equal to 20. Results for sample sizes between 20 and 50 are comparable but still differ by a significant amount. However, when the continuity correction is applied to Z-tests with n=21-50, the normal approximation and the exact binomial test results are virtually identical.

Power analyses conducted during the DQO indicated that a sample size of 26 would be sufficient when the uncorrected Z-test was used. However, when the exact binomial or the continuity corrected version of the Z-test was used, power analysis indicated a minimum sample size of 32. It was decided that 32 was the conservative choice; this number was increased to 34 to permit division of the year into 11-day sampling intervals. For illustrative purposes, both the exact

Table 5 2000 DISSOLVED OXYGEN MONITORING DATA

SAMPLE	DAY	DISSOLVED OXYGEN (MG/L)
1	02JAN	5.1
2	13JAN	4.5*
3	24JAN	10.3
4	04FEB	6.4
5	15FEB	8.2
6	26FEB	8.3
7	08MAR	7.3
8	19MAR	3.7*
9	30MAR	5.3
10	10APR	3.5*
11	21APR	5.1
12	02MAY	3.9*
13	13MAY	3.4*
14	24MAY	2.4*
15	04JUN	4.1*
16	15JUN	9.4
17	26JUN	6.2
18	07JUL	1.1*
19	18JUL	3.3*
20	29JUL	6.0
21	09AUG	5.0*
22	20AUG	3.2*
23	31AUG	8.5
24	11SEP	1.2*
25	22SEP	2.8*
26	03OCT	4.0*
27	14OCT	2.5*
28	25OCT	3.5*
29	05NOV	5.1
30	16NOV	7.2
31	27NOV	2.5*
32	08DEC	2.6*
33	19DEC	0.9*
34	30DEC	4.5*

binomial and Z-tests (with and without the continuity correction) well be used to test the null vs. the alternative hypothesis. Details of these tests are provided in Sections B2.4 and B2.5 of Appendix D.

Identify the Assumptions Underlying the Statistical Test. The model for the binomial distribution is appropriate for discrete response variables that meet to following assumptions:

1. The response can have only two outcomes (e.g., attainment, exceedance)
2. The underlying probability of exceedance, p, remains constant from sample to sample
3. Samples are obtained through an independent random sampling design

4. Verify the Assumptions of the Statistical Test

The exact binomial test is an example of a nonparametric test and as such does not require restrictive assumptions relating to the shape or the variability of the distribution. Thus no specific goodness of fit tests or graphical methods are needed to verify the assumptions. The three required assumptions can be verified by reviewing the sampling design and the data listing in Table 5. The response variable is clearly dichotomous. However, the second assumption is problematic. The binomial is most familiar as a model for the probability of obtaining various numbers of heads from repeated coin tosses. Assumption 2 essentially says that the binomial probability model requires that the very same coin be used for each toss. For example, if two different coins are used, one of which is fair (P=0.50) and other is not (p=0.75), the cumulative distribution function of the binomial will not provide correct estimates of the probabilities of obtaining specific numbers of heads. This assumption when applied to the DO sampling data requires that the exceedance probabilities at the monitoring site not change seasonally. Table 5 suggests that this assumption does not hold. Dissolved oxygen tends to be higher in colder months when less decomposition occurs. Thus the probability of exceedance is lower during the winter and early spring sampling periods than at other times of the year.

The assumption of independence would be valid if the assumption of constant-p were valid. However, because of the seasonality in exceedance probability, there is likely some autocorrelation in the data. In this case, the investigators were unable to make any specific adjustments or corrective actions to account for the violations of assumptions 2 and 3 , and simply proceeded to apply the exact binomial test to the data. In practice, when seasonality is observed in the data, a statistician experienced in the analysis of seasonal time series should be consulted to determine if an alternative approach or model should be employed.

5. Draw Conclusions from the Data

Perform the Statistical Hypothesis Test. The test statistic for the exact binomial distribution is just the number of samples out of 34 that had a DO value < 5.0 mg/l. As with all one-sample hypothesis tests we want to obtain a p-value that represents the probability of observing a test statistic greater than or equal to the one we have obtained from our data set. To do this, we need to know the expected distribution of the test statistic when the null hypothesis is true.

The probability distribution of the exact binomial for a population of 34 samples with an underlying 10% non-attainment probability is shown in Figure 18b. The histogram is the graphical representation of the discrete binomial probability density function. The total area within the darkened bars sums to 1.0. The probability that we desire is the blackened area that lies just to the right of the 20-sample point on the horizontal axis. This point is not shown in the figure because the there is no discernible probability of observing more than 9 exceedances out of 34. Table 6 lists the individual terms of the cumulative probability of the binomial distribution with n=34 and P=0.10. We can see that the probability of observing 13 or more exceedances is always <0.0001. In fact, the exact binomial probability of 20 or more non-attainment values from a population with only 10% non-attainment is 3.43×10^{-12}. Recall that we made a decision during the DQO process that if the probability of observing a value greater than or equal the sample test statistic was < 0.15 (i.e., we set α=0.15), we would reject the null hypothesis and conclude that the data did not support a decision that the water body attained the DO standard during 2000. Thus we must reject the null hypothesis that the true proportion of non-attainment at the monitoring site was ≤ 0.10.

The curve that delineates the normal approximation to the binomial with N=34 and p=0.10 is overlaid on the binomial probability distribution in Figure 18b. The total area under the bell-shaped curve is 1.0. The curve is smooth because it plots the probability distribution of a continuous, rather than a discrete, random variable. In this case, the random variable is Z:

$$z = \frac{Y - m}{s/n}$$

where $m = np$

y = the observed number of non-attainment values out of n samples

$s/n = \sqrt{np(1-p)}$

Substituting 34 for n and 0.10 for p, we get μ=3.4 and σ=1.75. Thus the Z for Y=20 non-attainment values is 9.49. This is an extremely large Z-score; 95% of the normal probability density lies between Z=± 2.0. Thus a value of Z as large as 9.49 has very little probability of occurring. In fact, its P-value is $< 10^{-25}$.

The continuity-corrected normal approximation for the upper-tailed, one-sided alternative hypothesis is calculated as:

$$z = \frac{Y - 0.50 - m}{s/n}$$

The result is Z=9.20 which likewise has $P < 10^{-25}$.

Fig. 18. (A) Frequency distribution of 34 area-adjusted DO means (mg/l) taken from the Mermantau River, at 11-day intervals, Jan-December 2000. (A) Relative frequency histogram and normal probability function ($\mu = 100.3$, $\sigma = 86.1$). (B) Expected frequency distribution of n=34 means under $H_{0:}$ the distribution of exceedances is normal with $\mu = 3.4$ and $\sigma = 1.75$.

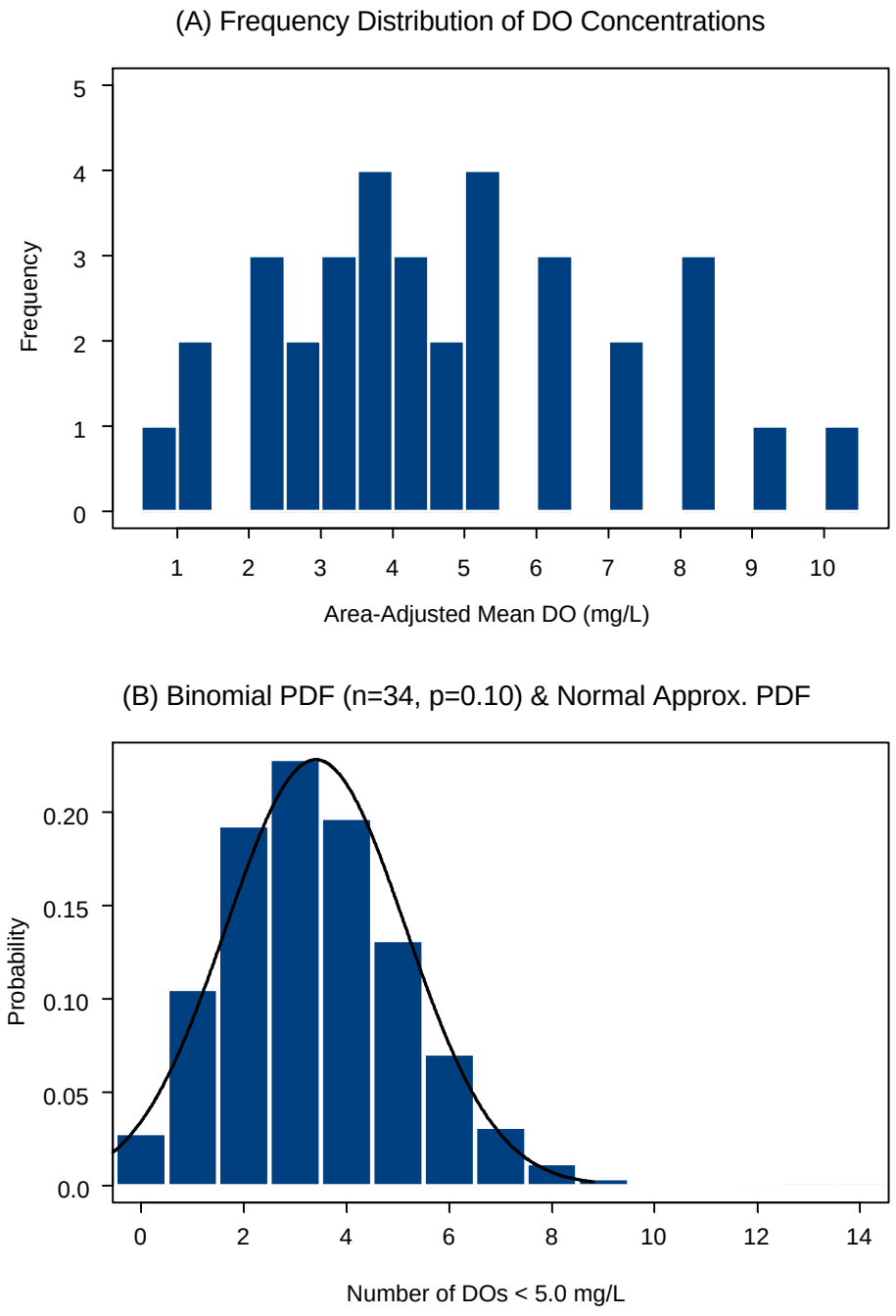


TABLE 6 CUMULATIVE PROBABILITIES OF 1-34 SUCCESSES
OUT OF N=34 TRIALS (P=0.10)

NUMBER SAMPLES WITH DO < 5.0 MG/L	CUMULATIVE BINOMIAL PROBABILITY
1	0.9722
2	0.8671
3	0.6745
4	0.4462
5	0.2496
6	0.1185
7	0.0481
8	0.0169
9	0.0051
10	0.0014
11	0.0003
12	<0.0001
13	<0.0001
14	<0.0001
15	<0.0001
16	<0.0001
17	<0.0001
18	<0.0001
19	<0.0001
20	<0.0001
21	<0.0001
22	<0.0001
23	<0.0001
24	<0.0001
25	<0.0001
26	<0.0001
27	<0.0001
28	<0.0001
29	<0.0001
30	<0.0001
31	<0.0001
32	<0.0001
33	<0.0001
34	<0.0001

The horizontal axis scale in Figure 18b is incremented by the number of observed non-attainments. This is the appropriate scale for the binomial distribution. The overlaid normal probability curve shows the probabilities associated with the z-transformation of the binomial scale values. For example, the value of 8 non-attainments on the binomial scale translates into (8-3.4)/1.75 = 2.63 on the Z-scale. The proportion of the normal curve to the right of 2.63 is the normal approximation of the area that lies to the right of 8 in the histogram of the exact binomial probability distribution. Similarly, the p-value of $<10^{-25}$ associated with the Z-score of 9.20 is an approximation of the exact binomial P-value of 3.43×10^{-12}. Both P-values lead to the rejection of H_0.

Draw Study Conclusions. Because the sample size (n=32) specified in the DQO insured that the false positive rate will be no greater than 0.15, we can be assured that the results, based on n=34, have an acceptably low probability of falsely rejecting the null hypothesis. In fact, the extremely large proportion of non-attainments (i.e., 0.59) places this estimate well beyond the upper boundary of the gray zone in Fig. 4 where the false negative rate is vanishingly small. Thus both the data (Table 5; Fig. 18 a) and the test statistics lead us to conclude that the dissolved oxygen content at the monitoring station was lower the 5.0 mg/l action level, more than 10% of the time during 2000. These results provide a basis for listing the reach of the river on which the hog operation is located and with the imposition of further corrective actions by the hog facility operators.

Evaluate Performance of the Sampling Design. Because this testing scenario involves a binomial proportion, the variability of the response was known exactly before the study was initiated. Thus it is not necessary to review the sampling design after the data were collected.

C.3.5. Data Quality Assessment Case History: Monitoring 3-Year Mean Turbidity in a River Reach

This example concerns the comparison of the sample mean value of a nonpriority water quality variable to its criterion value as set by state environmental regulations. The example is presented in a continuous format to show how to implement the five-step DQA process for a study that was not designed through the DQO process.

0. Background

Turbidity is the degree of opaqueness produced in water by suspended particulate matter. Turbidity may be quantified by measuring the degree to which light is scattered and/or absorbed by organic and inorganic material suspended in the water column. The standard unit of measure for turbidity is the nephelometric turbidity unit (NTU); larger NTU values indicate increased turbidity and decreased light penetration. The greater the amount of absorption and/or scattering, the less transparent the water, and the shallower the depth to which sunlight can penetrate. Because sunlight may only penetrate a few inches below the surface of waters with high turbidity, photosynthesis may become confined to the surface layers, leading to reductions in dissolved oxygen and productivity in deeper waters. Decreases in primary production associated with increases in sedimentation and turbidity may produce negative cascading effects through depleted food availability to zooplankton, insects, freshwater mollusks, and fish. Direct effects

at each trophic level include mortality, reduced physiological function, and avoidance; however, decreases in available food at different trophic levels also result in depressed rates of growth, reproduction, and recruitment (Henley et al. 2000).

Water quality regulations in a fictitious parish in Louisiana required the 3-year mean turbidity of all lotic waters to be less than 150 NTU. Beginning in January 1980, the parish began to collect data on turbidity and other water quality parameters in a 20-mile reach of the Mermentau River. During this 20-year period, turbidity was recorded *in situ* using several different models of tubidimeters. The monthly turbidity data (sorted by NTU value) for 1980 - 2000 are shown in Table 3. The subset of the data for the 3-year period between 1997 and 1999 is shown in Table 1 of Appendix D. The sampling schedule was not designed through the application of a DQO process; rather it was based on a combination of budgetary considerations and past practices that were designed to capture seasonal variability.

1. Review the Study DQOs and Associated Sampling Design

Review the Study Objectives. The basic research problem was to estimate a 3-year mean turbidity value for a 20-mile reach of the Mermentau River and to determine whether, given the sampling error associated with the underlying sampling design, the 3-year mean was less than the state water quality standard of 150 NTU. The sampling design called for 21 turbidity measurements taken at 1-mile intervals along the river reach. Measurements were taken such that 10 readings were taken 10 meters from one shore and 11 were taken 10 meters from the opposite shore in a systematically alternating pattern. The near shoreline (i.e., left or right) chosen for the first measurement was alternated monthly. Operation and deployment of an YSI-6026 wiped turbidity sensor were made following the manufacturer's guidelines. Data were collected between 9:00 AM and 12:00 PM on the 15th day of each month or on the nearest Friday or Monday when the 15th fell on a Saturday or a Sunday. All scheduled collections were made; there are no missing data [Table 1 (Appendix D)]. The unit of analysis was the monthly mean turbidity (NTU) computed from each month's systematic sample of 21 turbidity measurements.

Translate the Objectives into Statistical Hypotheses: Because the parish regulations were written in terms of 3-year mean turbidity values, the population parameter of interest was the mean turbidity value of the specific 20-mile reach of the Mermentau River. Parish water quality standards required this mean to be less than 150 NTU. Therefore the hypotheses of interest are, "3-year mean turbidity $\geq$ 150 NTU" and "3-year mean turbidity < 150 NTU".

There are two possible decision errors: 1) to decide that the turbidity in the river reach exceeds the criterion value when in fact it does not; or 2) to decide that the turbidity in the river reach is less than the criterion value, when in fact it exceeds it. Because of the profound ecological and economic risks associated with sustained high turbidity (i.e., fish kills, undesirable algal blooms, proliferation of undesirable and possibly pathogenic microorganisms, poor visibility for navigation, fouling of industrial intake pipes, etc.), the consequences associated with a decision error of the second type were deemed to be far more serious than those of the first type. Accordingly, the null hypothesis was that the turbidity in the river reach *exceeded* the criterion value ("3-year mean turbidity $\geq$ 150 NTU") and the alternative hypothesis was that the mean turbidity attained the state water quality standards ("3-year mean turbidity < 150 NTU"). (See

Section C.3.3 for more information on choosing the most appropriate null and alternative hypotheses.)

Develop Limits on the Decision Errors: The gray region associated with an alternative hypothesis of the form, "3-year mean turbidity < 150 NTU" is bounded on the right by the action level (150 NTU) on an the left by a value that is less than the action level. Values of the 3-year mean that fall within the gray region are subject to high probabilities of falsely declaring a population 3-year turbidity mean that is < 150 NTU, to be > 150 NTU; this is called a "false-negative" decision error. The consequences of this type of error include unnecessary expenditure of resources on remedial actions, unnecessary regulatory action, and unnecessary economic hardship on individuals and companies that would be affected by erroneously listing the river and restricting its uses. Therefore we would like the gray region to be as narrow as possible. However, reducing the width of the gray region will require larger sample sizes (and hence greater expense) if the false negative error rate is to be maintained at a reasonably small value. With these cost-benefit relationships in mind, the lower bound on the gray region initially was set at 120 NTU (20% below the action level). It was decided that this value should be subject to revision if, after setting the minimum acceptable decision error rates, it required a sample size that was larger than the 36 monthly means that were available to make the decision.

Because the potential economic and social costs of falsely rejecting the null hypothesis were high, it was decided that the two types of decision error rates should be set equal to one another. Moreover, because the maximum sample size was already fixed at 36 and there was a desire to keep the width of the gray region as small as possible, it was decided that moderate sized error rates were all that was feasible. Therefore the false acceptance and false rejection error rates were both set to 0.15. Using the USEPA DEFT software with these specifications, it was found that a minimum of 63 samples would be required. Consequently, the lower bound of the gray region was extended to 110 NTU with the result that the DEFT software computed a new minimum sample size of 31. Thus, the given sample size of 36 insured a false negative error rate of ≤ 0.15 and a false positive rate of 0.15 with a gray region extending from 110 NTU – 150 NTU. These relationships are illustrated for the log-scale turbidity data in Figure 19.

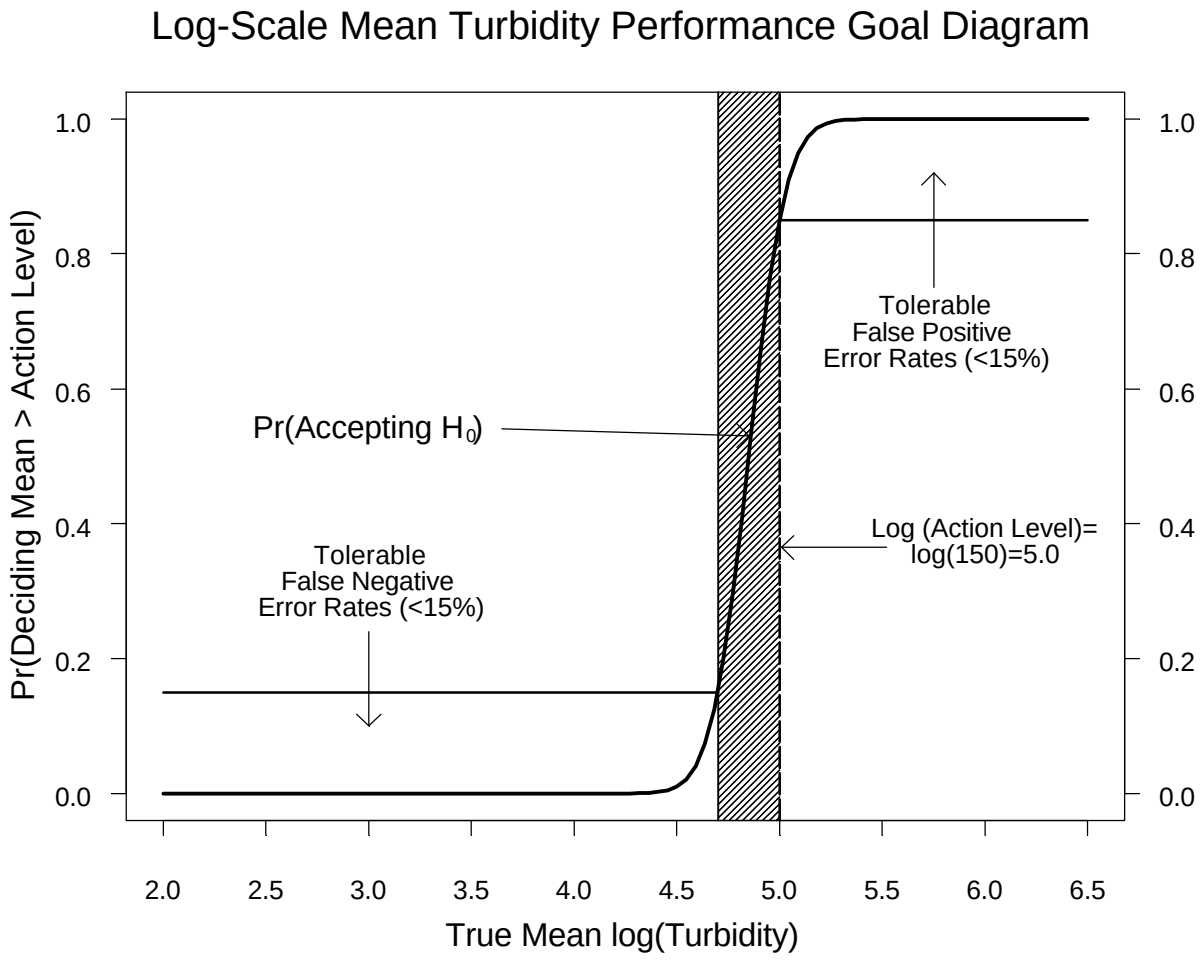
2. Conduct a preliminary Review of the Data

The 36 monthly mean turbidity values in NTU and in units of the natural log of the NTU values are shown in Table 1 (Appendix D). Previous analyses of these data revealed that they were lognormally distributed; thus, all of the calculations associated with the retrospective DQO steps have been based on consideration of the log-scale turbidity data. By inspection, there is clearly seasonality in the turbidity data and it appears that measurements taken in consecutive months are more alike than those taken farther apart in time. This suggests that some adjustment for seasonality and temporal autocorrelation may be required when the data are statistically analyzed.

3. Select the Statistical Test

Selecting the Test. Although the water quality criteria for turbidity were written in terms of a 3-year mean, we will make the assumption that the regulation was meant to apply to the center of

Fig. 19. Decision performance goal diagram for a one-sample t-test of H_0: population mean = 5.0 vs. H_a: population mean < 4.7, when the standard deviation =250, the sample size=30 and the width of the gray region is 0.30 (minimum detectable effect size=d=0.30 on the log-scale).



the distribution of the turbidity values in the target population. As explained in section C.2.2 (pages 38-40), the mean is interpretable as a measure of central tendency (i.e., as the median) when the data are normally or at least symmetrically distributed. Twenty years of monitoring in the Mermentau River and historical data from other nearby rivers had shown that monthly turbidity in these waters was consistently lognormally distributed. The back-transform of a mean computed on the log-scale is called the geometric mean. The geometric mean of lognormal data is an unbiased estimator of the population median on the original scale (Section C.2.2). Thus a statistical test that compares the log-scale mean turbidity to the natural log of the action value is equivalent to comparing the median on the original scale to the action level (i.e., 150 NTU). The following pair of hypotheses is appropriate for such a comparison:

$$H_0: \text{mean of } \log(\text{ Turbidity}) \geq \log(150)$$
$$H_a: \text{mean of } \log(\text{Turbidity}) < \log(150)$$

The appropriate statistical test for these hypotheses is the one-sample t-test against a lower one-sided alternative (see Box 8 and Section D.2). These were the hypotheses of interest for the turbidity assessment and therefore the one-sided t-test was used to support the decision of attainment/nonattainment of the 150 NTU criterion by the 1998-1999 Mermentau turbidity data.

The presumption behind this approach is that the regulation for a maximum mean turbidity of 150 NTU was based on the equivalence of the mean and the median in normal distributions. However, if the intention of the regulators was in fact to focus on the mean regardless of whether it was located in the center of the distribution, then different estimation and testing procedures would be required. Chen's test can be used to test the above hypotheses on the means of the untransformed turbidity data and specialized estimation techniques are available for forming confidence intervals around the untransformed mean (Millard and Neerchal 2001).

Identify the Assumptions Underlying the Statistical Test. The one-sample t-test for lognormal data is appropriate for continuous response variables that meet to following assumptions:

a. The distribution of the natural logarithms of the data values are approximately normal
b. The data values come from an independent random sample of the target population

4. Verify the Assumptions of the Statistical Test

Assessment of the distributional form of the turbidity data was described in detail in Section C.2.2. Q-Q plots were used to verify that the log-transformed turbidity was approximately normally distributed (Fig. 6). In Section C.1.7, the independence assumption was shown not to hold for the 1997-1999 turbidity data; significant seasonality and temporal autocorrelation were demonstrated. However, a modification to the usual t-test calculation was applied to provide a means of testing the hypotheses of interest. The modification employed an adjusted estimate of the sample variance and a formula for adjusted degrees of freedom for the t-test. The adjusted variance takes account of the estimated autocorrelation in the monthly time series data. Because autocorrelated measurements tend to be alike, the 36 monthly turbidity measurements contain redundant information. This is reflected in the adjusted degrees of freedom that were computed for the t-test: df=8 (Appendix D, Section D.2). Because the degrees of freedom for the one-sample t-test = n-1, the effective sample size is only n=9 or 25% of the actual sample size. This creates a situation in which the specified maximum allowable false negative error rate of 0.15 is

likely to be exceeded; i.e., the presence of significant autocorrelation decreases the effective samples size and the power of the statistical test.

5. Draw Conclusions from the Data

Perform the Statistical Hypothesis Test. The test statistic for the adjusted t-test is:

$$t_{a,df} = \frac{\bar{y} - k}{\frac{\sqrt{s^2_{adj}}}{\sqrt{nm}} \times \sqrt{\frac{1+\hat{f}_1}{1-\hat{f}}}}$$

where, $\bar{y}$ = the mean log-turbidity

k=log(150 NTU)

$t_{a,df}$ = the adjusted t-statistic with df degrees of freedom

n = the number of years of data

m = the number of months in a year

df = ((n×m)-m)/3

s^2_{adj} = the variance of the seasonally adjusted log-turbidity

$\hat{f}_1$ = the estimated autocorrelation in the seasonally adjusted log-turbidity at lag1

There were 12 monthly means in each of the three years of data, so df=8. The denominator of the above expression is the standard error of the 36-month turbidity time series. This expression is explained in Section D.2 of Appendix D; its value for the 1997-1999 data was 0.164 and the log-scale mean was 4.12, yielding an adjusted t-statistic of 5.43 with a P-value of 0.0003 for the lower one-sided alternative hypothesis that the true log-scale mean turbidity is < log(150) NTU. Because we have specified a false positive error rate of α=0.15 and the observed P-value is < α, we should reject the null hypothesis and accept the alternative hypothesis that the population 3-year median turbidity is < the 150 NTU action level. The actual false negative error rate associated with the effective sample size of 9 is β=<0.0001, substantially smaller than the planned rate of 0.15 that was associated with n=31.

Draw Study Conclusions. Because of the autocorrelation in the monthly turbidity measures, the effective sample size was only n=9, far below the minimum of n=31 that was specified in the DQO. This result demonstrates that even when the DQO steps are followed, if the independence assumption is violated, the estimated minimum sample sizes may be too small to maintain the desired decision error rates. In this case the geometric mean turbidity of the sample (61.6 NTU) was so far below the 150 NTU action value that even a sample size of n=9 had more than sufficient power to distinguish it from the action value. However, if the sample geometric mean had been closer to the action value, the false negative error rate (β) could easily have exceeded the value of 0.15 that was specified in the DQO. It is important to understand that the DQO estimates are contingent on the assumption that the conditions under which the DQO was carried out will prevail when the actual data are collected from the field. In this case that was not so, but as it turned out (rather fortuitously) the direction of the effects of these unforeseen conditions were favorable to maintaining decision error rates below the limits that were specified in the DQO.

Evaluate Performance of the Sampling Design. The results show that because of the substantial autocorrelation in the data, the sampling design employed was not adequate to provide data for a one-sided t-test with a specified false negative error rate ≤0.15. Unfortunately, restriction of the design to a 3-year period places severe limitations on what can be done. The decline of the effective sample size from 36 to 9 is a consequence of the seasonality and the autocorrelation in the turbidity time series. The usual remedy for inadequate sample size is to increase the number of samples that are collected, but in the case of the 3-year time series, this can only be done by increasing the sampling frequency within the 3-year period. This requires us to collect samples closer together in time, which will inevitably increase the autocorrelation. As the autocorrelation increases, the denominator term in the adjusted t-statistic shown above will get larger and the t-statistic will get smaller. Intuitively we can see that collecting more autocorrelated data, more frequently will not actually provide us with more useful information. Because of the redundancy in the data due to the autocorrelation, two measurements taken close together will simply be telling the same thing twice.

Ecologists and others who have analyzed annually fluctuating, seasonal environmental time series data have long recognized that decisions based on estimates and statistical tests of short-term environmental data (e.g., 3-year studies) may be prone to unacceptably high error rates. The only solution to the problem is to extend the duration of the study; some studies may extend for decades before it becomes possible to make good inferences from them. Generally speaking, once enough samples have been collected to estimate adequately the within-year seasonal effects (e.g., 12 monthly samples per year), the precision and decision error rates cannot be changed appreciably by simply increasing the number of samples within years. Thus, in cases where seasonality effects are pronounced and autocorrelation is high, it may not be possible to implement a DQO that will yield a sampling design that can ensure that specified decision error rates will be achieved within a 3-year period.

C.3.6. Data Quality Assessment Case History: Evaluating the 3-year Acute Maximum Concentration Criterion for Cadmium in a River Reach

This example concerns the comparison of the daily mean values of a priority toxic pollutant to its 3-year acute Criterion Maximum Concentration value (CMC) as set by EPA. The example is presented in a continuous format to show how to implement the five-step DQA process for a study that was not designed through the DQO process.

0. Background

Cadmium is a heavy metal whose chemical properties are similar to zinc. Cadmium does not occur uncombined in nature and is usually obtained as a byproduct of smelting and refining ores of zinc and lead. Cadmium is used principally for its anticorrosive properties in the electroplating of steel, in its sulfide form in the manufacture of paint pigments, and in the manufacture of batteries and other electrical components. Cadmium also occurs as a byproduct in many chemical fertilizers that are produced from phosphate ores. Cadmium enters the ambient air primarily from local smelting operations, it enters soil from local mining operations and from chemical fertilizers and it enters water from fertilizer runoff and/or industrial wastewater. Once in the water column, cadmium is rapidly absorbed by suspended particulates, which eventually settle into the sediments. From the sediments, it enters aquatic food chains where it tends to bioconcentrate in plants, crustaceans, and mollusks. When ingested by mammals, cadmium becomes concentrated in the liver and the kidneys.

Toxicity in mammals (including man) is due principally to kidney damage, which upsets calcium regulation and results in a net loss of calcium through the urine. Ultimately, this may lead to calcium depletion in bones and egg shells, which in turn may cause injury, reproductive failure, or death.

Mean concentrations of cadmium in unpolluted waters are typically < 1.0 μg/L. EPA has set the Criteria Maximum Concentration (CMC) of cadmium at 4.3 μg/L. The CMC, determined experimentally in laboratory studies, represents the highest concentration of a pollutant to which 95% of the freshwater aquatic organisms in an ecosystem can be exposed without suffering deleterious effects (Nowell and Resek 1994). If only one sample is taken in a 24-hour period, the pollutant concentration in that sample is the best estimate of the maximum concentration of the pollutant on that day, at that site. This value may be compared to the EPA CMC value to determine if the site attains the EPA standard on that day. However, EPA's CMC standards are written in terms of 3-year sampling periods. Specifically, the standard states that no more than one daily mean concentration may exceed the CMC for any given 3-year period at a single site. Theoretically, an aquatic ecosystem can recover from a single pollution event that exceeds the CMC during a 3-year period, but not from 2 or more such events. Thus, in practice, the total number of CMC exceedances for a specific pollutant (e.g., cadmium) in a 3-year period must be compared to the EPA standard of only one allowable CMC exceedance per 3-year monitoring period.

A fictitious parish in Louisiana routinely collects monthly samples of water from several lotic and lentic systems within the state. Concentrations of several pollutants are estimated from these samples and entered into an electronic database. Cadmium data collected during 1981-1983 from a reach of the Tangipahoa River in Southeastern Louisiana will be used in this case study to illustrate application of DQA procedures to the assessment of the once-in-3-years criterion for exceedances of the cadmium CMC. Because the monitoring program was designed prior to the development of EPA's DQO procedures, DQO principles will be applied retrospectively to frame the research question, formulate testable hypotheses, select an appropriate statistical test and establish bounds on the acceptable decision error rates.

1. Review the Study DQOs and Associated Sampling Design

Review the Study Objectives. The basic research problem was to determine whether, given the sampling error associated with the underlying sampling design, there was more than 1 daily exceedance of the cadmium CMC criterion (4.3 μg/L) in a reach of the Tangipahoa River, from January 1981- December 1983. Because the EPA CMC criterion for cadmium is written in terms of the number of days during a 3-year period that are allowed to have a daily mean cadmium concentration > the CMC, the target population is actually the 1095 days in the 3year period, each of which potentially could have had a daily mean cadmium concentration > 4.3 μg/L. Thus the primary objective of the study is to estimate the total number of days or, equivalently, the proportion of days during the three year period when the daily mean cadmium concentration was > 4.3 μg/L.

The sampling design called for one cadmium concentration measurement to be taken from a single sampling station in the 5-mile river reach within the first 3 days of each month during 1981-1983. A 200 ml sample of water collected from the monitoring station was filtered at the collection site. The filter with particulates and the filtrate were returned to the laboratory and subjected to further chemical analysis. In the laboratory, both particulates and filtrate were

digested using 55% nitric and 70% perchloric acids in the ratio of 2:1. Cadmium concentrations were determined by means of flame atomic absorption spectrophotometry. Accuracy of each monthly measurement was insured by comparison to known analytical standards. The unit of analysis was the value of the mean cadmium concentration (μg/L) at the Tangipahoa monitoring station, on one day.

Translate the Objectives into Statistical Hypotheses: In order to answer the question of interest, "Was there more than 1 day during 1981-1983 when the mean daily cadmium concentration was greater than 4.3 μg/L?", we need an estimate of the total number of days with exceedant cadmium concentrations. The criterion for the cadmium CMC says that no more that 1 exceedance of the daily CMC can occur within a 3-year period. This leads naturally to the following pair of hypotheses:

H_0: total number of exceedant days ≤ 1
H_a: total number of exceedant days > 1

However, since the total number of days in a 3-year period is fixed (N=1095), the hypotheses could be framed in terms of the proportion of exceedant days:

H_0: proportion of exceedant days ≤ 1/1095 ≤ 0.0009
H_a: proportion of exceedant days > 1/1095 > 0.0009

The same probability model, the hypergeometric, can be used to evaluate both pairs of hypotheses (See Appendix D, Section D.6; Buonaccorsi 1987; Wendell and Schmee 2001). Like the binomial, the hypergeometric distribution model is appropriate for binary outcomes; e.g., attained vs., exceeded. The hypergeometric differs from the binomial model in that it assumes a finite population. For example, in the case of the CMC criterion, we are attempting to infer from a sample of 36 days to a "population" that contains exactly 1095 days. This contrasts with the use of the binomial distribution to infer from a small number of water samples to a river containing an essentially infinite number of such samples (e.g., see dissolved oxygen DQA case history).

For convenience, we will frame all statistical hypothesis tests and their associated error rates in terms of the proportion of exceedant days in a 3-year monitoring period. However, the hypergeometric distribution could easily be employed to test the equivalent hypotheses about the total number of exceedant days that occurred during the 3-year period. Given the second pair of hypotheses (above), there are two possible decision errors: 1) to decide that the proportion of daily mean cadmium concentrations >4.3 μg/L, exceeds 1/1095 when in fact it does not; or 2) to decide that the proportion of daily mean cadmium concentrations > 4.3 μg/L is less than or equal to 1/1095, when in fact it is not. Although the ecological and public health consequences of sustained high cadmium concentrations may be severe, the local economic and social consequences of incorrectly declaring the river reach to be in violation of the CMC once –in-3-years standard were considered more deserving of protection. Accordingly, the null hypothesis was that the proportion of exceedant days during the 1981-1983 monitoring period was ≤ 1/1095 and the alternative hypothesis was that the proportion of exceedant days was > 1/1095 (i.e., H_a: the mean daily cadmium concentration exceeded the CMC standard of 4.3 μg/L on more than 1 day during 1981-1983). (See Section C.3.3 for more information on choosing the most appropriate null and alternative hypotheses.)

Develop Limits on the Decision Errors: The gray region associated with an alternative hypothesis of the form, "The proportion of days during 1981-1983 with a mean cadmium concentration > 4.3 µg/L is > 1/1095" is bounded on the left by the action level (1/1095) on the right by a value that is greater than the action level (Fig. 20). Values of the proportion of exceedant days that fall within the gray region are subject to high probabilities of falsely declaring that the 3-year exceedance was < 1/1095 when in fact it was > 1/1095; this is called a "false-negative" decision error. The consequences of this type of error include possibly severe ecologic and pubic health effects. On the other hand EPA feels strongly that 2 or more exceedances in a year (i.e., ≥ 2/1095) should be cause for TMDL listing of the affected water body. Therefore it was decided that the gray region should be bounded on the left by 1/1095 and on the right by a proportion slightly less than 2/1095. Because 2/1095 = 0.00182, the upper bound was set at 0.0018. This means that the proportion associated with two exceedances will lie just outside of the gray and thus not be subject to excessively high false negative decision error rates.

Because the potential economic and social costs of falsely rejecting the null hypothesis were high, it was decided to set the Type I error rate at 0.05. The ecologic and public health consequences were deemed to be less important than the economic consequences so a higher Type II error rate of 0.20 was specified. The USEPA DEFT software does not compute samples sizes and power associated with hypergeometric probabilities. However these can be computed with other commercially available software (e.g., PASS). Specifying a lower bound of 1/1095 (0.00091) and an upper bound of 0.0018 for the gray region, a Type I error rate of 0.05, a Type II error rate of 0.20, and population size of 1095 days, PASS indicated that a minimum of 980 days out of 1095 would need to be sampled in order to meet the specifications. The reason for this large sample size is the extremely narrow gray region (width=0.0009); however, given the EPA standards which require listing for the occurrence of 2 or more exceedant days out 1095, there is no other acceptable choice for the gray region.

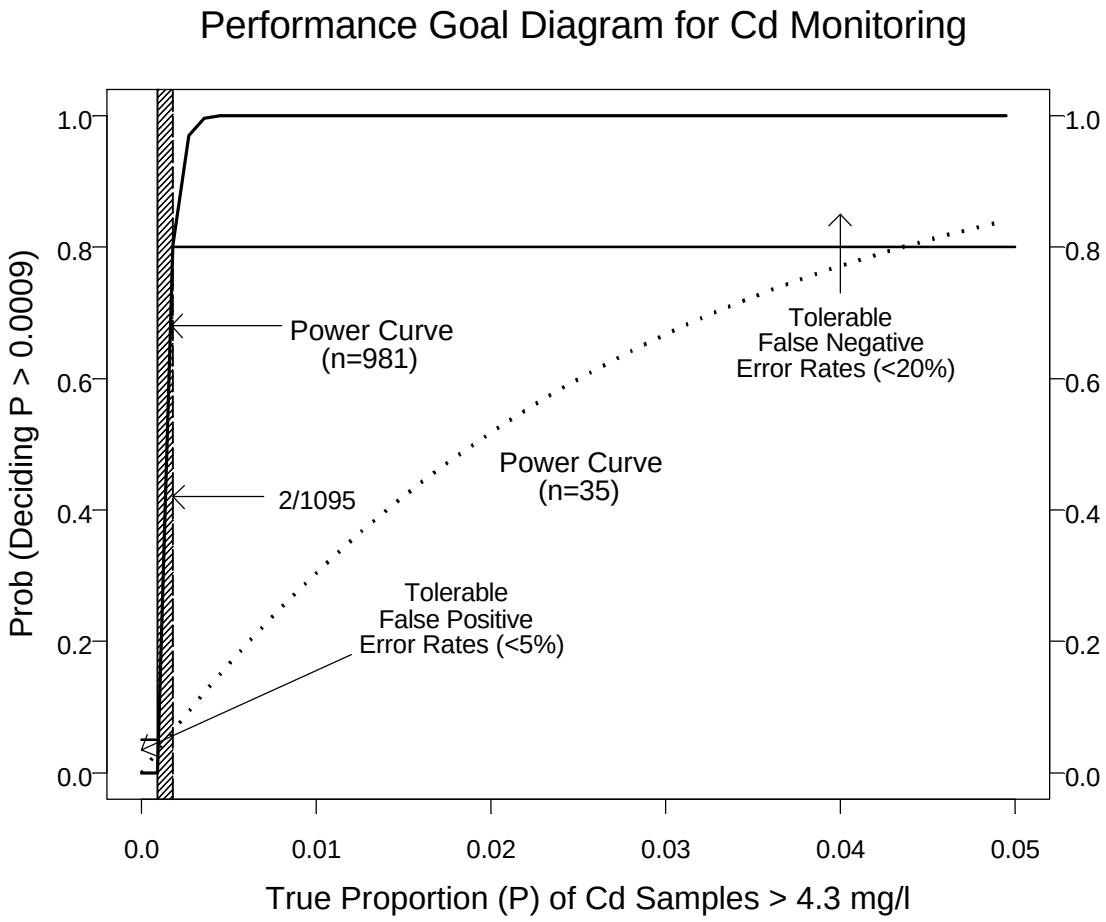
2. Conduct a preliminary Review of the Data

Only 35 of the scheduled 36 monthly sample measurements were available for analysis; the November 1982 sample was lost in transit to the laboratory. The 35 monthly cadmium concentration values are shown in Table 7. Although the two highest cadmium concentrations occurred in February, there is no clear pattern of seasonality or autocorrelation in the 1981-1983 cadmium data. Of the 35 concentrations, only one (5.0 µg/L in February 1981) exceeded the cadmium CMC. This suggests that the best estimate of the proportion of exceedant days during the 19981-1983 sampling period is the sample estimate, 1/35= 0.0286, a value that is clearly larger than the EPA standard of 0.0009 (1/1095). However, the point estimate does not account for the uncertainly due to the sampling. We will need to use the hypergeometric model in order to address the uncertainty in the sample estimate.

3. Select the Statistical Test

Selecting the Test. As noted above, the hypergeometric distribution is the appropriate probability model for binary outcomes from a finite population (i.e., a population of known fixed size, N). Although a one-sample z-test based on the normal approximation to the hypergeometric distribution is available (Cochran 1977), it does not perform well for sample sizes < 400 or when the population proportion is small (e.g., 1/1095; Johnson and Kotz 1969). Thus a decision was made to use the exact hypergeometric test to provide support for H_0 vs. H_a.

Fig. 20. Decision performance goal diagram for an exact hypergeometric test of H_0: the population 3-year exceedance rate is = 1/1095 vs. H_a: population mean > 1/1095, when the sample size=980 daily samples (solid power curve) or the sample size=35 daily samples (dotted line), and the width of the gray region is 1/1095 (minimum detectable effect size=d=1 daily exceedance).



If the upper tailed exact probability of observing a proportion larger than the sample proportion, is less than α (i.e., 0.05), the null hypothesis will be rejected and the population will be considered to have had more than one exceedant day during the monitoring period. Otherwise, the null hypothesis that the rate of exceedance is = 0.0009 (i.e., that there was not more than one exceedance in the river reach between Jan 1, 1981 and December 31, 1983) will be accepted.

Identify the Assumptions Underlying the Statistical Test. The Exact hypergeometric test is appropriate for a binary response (e.g., attains vs. exceeds) expressed as a proportion (e.g., the proportion of exceedances) or as a total count (e.g., total number of exceedances). The assumptions required for the exact hypergeometric test are:

1. The response can have only two outcomes (e.g., attainment, exceedance)
2. The underlying probability of exceedance, p, remains constant from sample to sample
3. Samples are obtained through an independent random sampling design.

4. Verify the Assumptions of the Statistical Test

The exact hypergeometric test is an example of a nonparametric test and as such does not require restrictive assumptions relating to the shape or the variability of the distribution. Thus no specific goodness of fit tests or graphical methods are needed to verify the assumptions. The three required assumptions can be verified by reviewing the sampling design and the data listing in Table 1. The response variable is clearly dichotomous. The second assumption would appear to be true for these data, at least there is no evidence in the Table 7 to suggest that the probability of exceedance deviates very much from zero over the course of the 3-year sampling period. Because the proportion of exceedances appears to be constant and near zero at all times, there is no suggestion that measurements taken close together in time are more likely to be exceedant than measurements taken at longer intervals; thus, the independence assumption appears to hold for these data.

5. Draw Conclusions from the Data

Perform the Statistical Hypothesis Test. The upper tailed hypergeometric probability was computed as 1-the cumulative hypergeometric probability function evaluated with r=1 (p=1/35), N=1095, n=35, and k=1 (P_0=1/1095). These values were substituted into Equation 4 of Appendix D to yield a cumulative probability of 1.0; thus the upper tailed p-value was zero. Since a=0.05 was specified, and the p-value was less than 0.05, the sample data did not support the null hypothesis that there was = 1 exceedant day in the river reach during 1981-1983. In fact, the corresponding estimate of the total number of exceedant days that occurred during 1981-1983 was 32 with a lower 1–sided 95% confidence limit of 2. Thus given the sampling error associated with the sample of n=35 days, we have 95% confidence that there were at least 2 days during 1981-1983 when the daily mean cadmium concentration exceeded the CMC. However, the false negative probability associated with this test is 0.937; i.e., in a 3-year period that actually had exactly 2 exceedant days, the probability that the sample estimate of the cumulative hypergeometric probability based on a sample of 35 days would lead to erroneous acceptance of the null hypothesis is 0.937. In plain English, this result says that if 1 exceedant day is found in a random sample of 35 days, there is a 95% probability that there was at least 1 additional day

TABLE 7 THE 1981-1984 CADMIUM DATA FROM THER TANGIPAHOA RIVER, LA

YEAR	MONTH	CADMIUM ug/L
1981	1	0.9
	2	5.0*
	3	3.7
	4	1.3
	5	2.7
	6	2.0
	7	2.7
	8	2.7
	9	2.4
	10	3.2
	11	2.1
	12	1.4
1982	1	2.0
	2	1.1
	3	1.8
	4	2.8
	5	1.1
	6	2.1
	7	1.6
	8	0.5
	9	1.2
	10	1.9
	12	0.6
1983	1	0.1
	2	3.9
	3	0.8
	4	0.2
	5	0.8
	6	1.1
	7	1.5
	8	0.8
	9	0.6
	10	0.2
	11	0.5
	12	0.4

* Concentration Exceeds the 4.3 mg/l Standard

that was exceedant during the 3-year period. Moreover, the high false negative rate tells us that unless the number of exceedant days (out of 1095) is much greater than 2, the exact hypergeometric probability estimate is not likely to indicate non-attainment. The null hypothesis was rejected in this case only because one exceedance in 35 days implies that there was likely a large number of exceedant days between January 1 1981 and December 31, 1983.

Draw Study Conclusions. The results of the exact hypergeometric test indicate a zero probability that the baseline assumption that there was ≤ 1 exceedant days in three years was true. This result indicates that that a serious problem with cadmium pollution likely occurred during 1983-1981. However, the very low power (0.063) associated with the test for a sample size of 35 is cause for concern for future studies in which the 1-in-3-years-exceedence criterion is to be based on similar sized samples.

Evaluate Performance of the Sampling Design. The low power associated with the once-per-month sampling design suggests that much larger sample sizes are required to protect against falsely concluding that there were less than 2 CMC exceedances in a 3-year period. In fact a power analysis based on the exact hypergeometric distributions shows that if we desire =0.05 and a false negative rate of 0.20, when there are as few as 2 exceedant days out of 1095 the minimum required sample size is 980 days; i.e., we would need to sample nearly every day to have an 80% chance of distinguishing whether there is more than one exceedance in a 3-year period. These relationships are illustrated in Fig. 20, which shows the associated power curves for the actual sample size (n=35 days) and the minimum sample size required to meet the DQOs (n=980). When n=980, and two exceedances actually occur within the 3-year period, the exact hypergeometric test has 80% power. However, when the sample size is 35 days a power of 80% is not achieved until 4.41% of the 1095 days (i.e., 49 days) are exceedant. Thus although EPA regulations require the listing of any water body with 2 or more exceedances, the sample size used for this study is reliable only for detecting exceedances that occur at a rate of 49 or more per 3 years of monitoring.

C.4. References

Browne, R.H. 1995. On the use of a pilot sample for sample size determination. Statistics in Medicine. 14:1933-1940.

Cleveland, W.S. 1993. Visualizing Data. Hobart Press. Summit, New Jersey, USA.

Elliot, J.M. 1977. Some methods for the statistical analysis of samples of benthic invertebrates. Freshwater Biological Assoc. Ambliside, Cumbria UK

Hengeveld, R. 1979. The analysis of spatial patterns of some ground beetles (Carabidae). Pages 333-346 *in:* M. Cormack and J.K. Ord, editors, Spatial and temporal analysis in ecology. International Co-operative Publishing House, Fairfield, Maryland, USA.

Henley, W.F., M.A. Patterson, R.J. Neves and A.D. Lemly. 2000. Effects of sedimentation and turbidity on lotic food webs: A concise review for natural resource managers. Reviews in Fisheries Science. 8(2):125-139

Millard, S. P. and N. K. Neerchal. 2000. Environmental Statistics with S-PLUS. CRC Press, Boca Raton, FL.

Nowell, L. and E.A. Resek. 1994. Summary of National Standards and Guidelines for Pesticides in Water, Bed Sediment, and Aquatic Organisms and their Application to Water-Quality Assessments. UGSS (http://ca.water.usgs.gov/pnsp/guide/guide_6.html)

O'Brien, R.G. 1998. A tour of UnifyPow: a SAS module/macro for sample size analysis. Proceedings of the 23rd SUGI Conference, 1346-1355. SAS Institute, Cary, NC.

Ord, J.K. 1979. Time series and spatial patterns in ecology. Pages 1-94 *in:* M. Cormack and J.K. Ord, editors, Spatial and temporal analysis in ecology. International Co-operative Publishing House, Fairfield, Maryland, USA.

Peterson, S.A., N.S. Urquhart, and E.B. Welch. 1999. Sample representativeness: a must for reliable regional lake condition estimates. Environ. Sci. Technol. 33:1559-1565.

Steel, G.D., J.H. Torrie and D.A. Dickey. 1996. Principles and procedures of statistics: a biometrical approach. McGraw-Hill, New York.

Stephan, C.E., D.I. Mount, D.J. Hansen, J.H. Gentile, G.A. Chapman, and W.A. Brungs. 1985. Guidelines for Deriving Numerical National Water Quality Criteria for the Protection of Aquatic Organisms and their Uses: EPA.

Thompson, S.K. 1992. Sampling. J. Wiley and Sons. New York, New York.

USEPA. 1999. National recommended water quality criteria – Correction. EPA 822-Z-99-001.

C.5 Glossary

alternative hypothesis - In a statistical hypothesis test there are two competing hypothesis, one of which, the alternative hypothesis, describes the set of conditions complementary to those described under the null hypothesis. For example: if the null hypothesis states that the mean pH of a set of samples is less than or equal to 5.0, the alternative hypothesis must be that the mean pH is greater than 5.0.

ANOVA Model - an acronym for Analysis of Variance. ANOVA models are linear models in which the total variance in a response is partitioned into two components: one due to treatments (and possible interactions among them) and the other due to random variability. In the simplest case where there is only one treatment factor, if the treatments have no effect on the response, the ratio of the variance components should be close to 1.0. If the treatments effect the response means, the ratio of the treatment component to the random component will be greater than one. Under the null hypothesis that the treatments have no effect, the sampling distribution of the ratio of the two variance components, each divided by their respective **degrees of freedom**, will be an F-distribution.

ARIMA Model - an acronym for autoregressive integrated moving average model. ARIMA models are linear models for regression and/or discrete treatment effects, measured through time, on responses that have been differenced at the appropriate **lag** distances.

autocorrelation - the internal correlation of a set of measurements taken over time and/or space. The correlation arises from the fact that points closer together in space and/or time tend to be more alike than those that are further apart. The autocorrelation function (either spatial or temporal) is a mathematical expression that relates the strength of the correlation to the distance (called the **lag**) between measurements.

Bayesian statistical inference - An approach to **inference** or **estimation** in which a process (e.g., a random binomial process) is proposed for the generation of a set of data. A mathematical model called a **likelihood** is specified for the process, such that the model parameters are random variables. A distribution, called the prior distribution, is developed for the parameters based on what is known about them, prior to collection of the data. Data are then collected and a mathematical principle called Bayes theorem is used to derive a second distribution of the parameters, called the posterior distribution, from the data and the prior distribution. The appropriate inference is then obtained from the posterior distribution. The Bayesian approach differs from the the classical frequentist approach in that is utilizes the investigator's prior knowledge of the system through the prior distribution.

bias - the systematic or persistent distortion of a measurement process that causes errors in one direction.

binary characteristic - a characteristic that can only have two possible values.

census - a study that involves the observation and/or measurement of every member of a population.

confidence interval - a range of values, calculated from the sample observations, that is believed, with a particular probability, to contain the true population parameter value. For

example, a 95% confidence interval implies that if the estimation process were repeated again and again, then 95% of the calculated intervals would be expected to contain the true parameter value.

confidence level (also called the the confidence coefficient) - the probability that the confidence interval will include the true parameter value; Equivalently, 1-the probability (α)that the true value is *not* contained within the interval .

continuous random variable - A random variable which may take on an infinite number of values.

convenience sample - a sample collected from a target population without implementation of a probability-based design. Sampling units are selected based on ease of data collection, without clear reference to an underlying frame; e.g., the collection of water samples near bridges rather than randomly throughout the stream reach to which an inference is desired. Because many (perhaps the majority) of the population sampling units have little or no probability of selection to the sample and because sample coverage typically is restricted to some potentially homogeneous subset of the target population, data from convenience samples are not valid for statistical inference to the target population.

correlation coefficient – A scale-invariant measure of the association between 2 variables that takes on values between –1.0 and +1.0. The correlation coefficient has a value of plus one whenever an increase in the value of one variable is accompanied by an increase in the other, zero when there is no relationship (i.e., the 2 variables are independent of one another), and minus one (-1) when there is an exact inverse relationship between them.

correlogram - a plot or graph of the sample values of the autocorrelation coefficient of a time series against different values of the **lag**.

decision error - an error that occurs when data misleads an investigator into choosing the wrong response action, in the sense that a different action would have been taken if the investigator had access to unlimited "perfect data" or absolute truth. In a statistical test, decision errors are labeled as false rejection (Type I) or false acceptance (Type II) of a null hypothesis.

degrees of freedom (df) - As used in statistics, df has several interpretations. A sample of *n* values is said to have *n* degrees of freedom, but if *k* functions of the sample values are held constant, the number of degrees of freedom is reduced by *k*. In this case, the number of degrees of freedom is conceptually the number of independent observations in the sample, given that *k* functions are held constant. By extension, the distribution of a statistic based on *n* independent observations is said to have *n-p* degrees of freedom, where p is the number of parameters of the distribution.

discrete random variable - A random variable which may take on only a finite number of values.

dispersion - the amount by which a set of observations are spread out from their mean and/or median.

effect size - In a **one-sample test**, the difference between the sample mean and a pre-specified criterion or standard value. In a **two-sample test**, the effect size is the expected difference between the mean of a treatment group or ambient site vs. the mean of a control group or reference site. Associated statistical tests typically evaluate the null hypothesis of a zero effect size vs. the alternative that the effect size is nonzero.

effective sample size - When data are collected from cluster-correlated populations, there is redundancy in the information carried by more highly correlated individuals. Thus, correlated individuals carry less information than do uncorrelated individuals. The effective sample size is the number of uncorrelated individuals from a simple random sample that would carry information equivalent to the information in the sample of correlated individuals. The effective sample size is always less than the apparent sample size; how much less, is a function of the strength of the correlation and the sampling design that was used to collect the data.

estimation - the process of providing a numerical value for a population parameter on the basis of information collected from a sample.

experimental design - the arrangement or set of instructions used to randomize subjects to specific treatment or control groups in an experimental study. Such a procedure generally insures that results are not confounded with other factors and thus provides scientifically defensible inferences regarding causal effects of the treatments.

exploratory data analysis (EDA) - an approach to data analysis that may reveal structure and/or relationships among measured or observed variables in a data set. EDA emphasizes informal graphical procedures that typically are not based on prior assumptions about the structure of the data or on formal models.

extreme values - the largest and smallest values (and perhaps their neighboring values) among a sample of observations.

frequentist statistical inference - an approach to statistics based on the likelihood of an observed result in a large or infinite number of independent repetitions of the same sampling or experimental procedure (e.g., see the frequentist definition of the **confidence interval** in this glossary).

geometric mean - a measure of central tendency calculated by back-transforming the mean of a set of log-transformed observations. If the original data come from a log-normal distribution, the sample geometric mean will provide an unbiased estimate of the sample **median**.

heterogeneous - a term denoting inequality or dissimilarity of some quantity of interest (most often a variance) in a number of different groups, populations, etc.

homogeneous - a term denoting equality or similarity of some quantity of interest (most often a variance) in a number of different groups, populations, etc.

imprecision/precision - A term describing the degree of spread among successive estimates of a population parameter by a sample statistic. The standard error of a sample estimator (e.g., the standard error of the mean) is a measure of imprecision/precision in the estimator. A high degree of spread (imprecision) will lead to an increased likelihood of a decision error, while a reduction

in spread will lead to a corresponding reduction in the likelihood of a decision error. Generally, precision will be increased by increasing the sample size.

independence - essentially, two events are said to be independent if knowing the outcome of one tells us nothing about the outcome of the other. More formally, two events *A* and *B* are said to be independent if Probability (A and B) = Prob(A) × Prob(B).

inference - the process of drawing conclusions about a population on the basis of measurements or observations made on a sample of individuals from the population.

lag - the distance, in units of time or space, between two events or locations. For example, an event occurring at time $t+k$ $(k>0)$ is said to lag behind the event occurring at time t, by an amount of time equal to lag k.

likelihood - the probability of a set of observed data, given the value of some parameter or set of parameters associated with a model for the underlying process that is hypnotized to have generated the data. For example, if we obtain 9 heads in 10 tosses of a coin, the likelihood of observing this result, given that the coin is fair (i.e., the binomial parameter p=0.50), is approximately 0.0098.

log-transformation - a transformation on a variable, X, obtained as, Y=ln(X) or Y=ln(x+c), where c is a constant positive value (e.g., 1.0). This transformation is useful for normalizing continuous variables with skewed distributions and/or stabilizing the variance of a variable whose standard deviation increases as a function of its mean, prior to statistical analysis with tests (e.g., t-tests) that assume normality and/or variance homogeneity..

maximum likelihood - a procedure for estimating the value of a parameter(s) of a model of the underlying process that produced some particular set of observations, such that the resulting estimate maximizes the likelihood of the observed data. For example, the maximum likelihood estimate for the binomial parameter P, given an experiment in which one obtains 9 heads in 10 tosses is P=0.90. The likelihood of obtaining 9 heads given an underlying binomial process with P=0.90, is 0.3874. Note that the estimate P=0.90 leads to a much larger likelihood than an estimate of P=0.50 does (0.0098; see definition of **likelihood**). In fact there is no value of P that will yield a larger likelihood of obtaining 9 heads out of 10 tosses than the estimate P=0.90; thus, P=0.90 is the maximum likelihood estimator of P.

median - in a sample or a population, the median is the value of a random variable such that half of the sampling units have larger values and half have smaller values. When the population or sample size is 2N+1, the median is the value of the random variable associated with the $N+1^{th}$ ordered sampling unit; when the population or sample size is 2N, the median is average of random variable values of the sampling units with ranks N and N+1. If the population is normal the median will equal the mean. If the population is log-normal the median will equal the **geometric mean**.

Monte Carlo methods - methods for finding solutions to mathematical and statistical problems by simulation. Often used when the analytic solution of the problem is intractable, or when real data are difficult to obtain, or to evaluate the behavior of statistics or models under a variety of hypothetical conditions which may or may not be directly observable in nature.

nonparametric statistical methods (also called distribution-free methods) - Statistical techniques of **estimation** and **inference** are often based on the assumption of some underlying parametric process; for example, one that generates responses that are normally distributed. By contrast, nonparametric estimation and testing procedures do not depend on the existence of an underlying parametric process. Consequently, nonparametric techniques are valid under relatively general assumptions about the underlying population. Often such methods involve only the ranks of the observations rather than the observations themselves.

noncentral t-distribution - the expected distribution of the t statistic when the alternative hypothesis is true. This contrasts with central t-distribution (usually referred to simply as the "t-distribution") which is the expected distribution of the t-statistic when the null hypothesis is true. In general, the probability that an observed t-statistic comes from a non-central t-distribution will be large (e.g., $P>0.20$) when the probability of that it comes from a central t-distribution is low (e.g., $P<0.001$), and vice versa.

null hypothesis - a hypthesis about some presumed prevailing condition, usually associated with a statement of "no difference" or "no association" (see also **alternative hypothesis**).

one-sample tests - Statistical tests that evaluate the null hypothesis that there is no difference between a sample statistic (e.g., mean or proportion) and a fixed criterion or standard value.

parametric continuous distribution - the probability distribution of a continuous random variable, specified by a mathematical function of the population parameters; e.g., the normal distribution with parameters, μ and σ^2.

parametric statistical methods - tests and estimation procedures that depend on the complete specification of an underlying parametric probability distribution of the population from which the sample was drawn. The estimators and test statistics that are based on functions of the estimates of the population parameters under the assumed population distribution model (e.g. normal) are valid only if the assumed population model is valid. An example is the t-statistic which assumes an underlying normal population.

percentiles - the set of divisions of a set of data that produce exactly 100 equal parts in a series of values.

population - any finite or infinite collection of "units" that completely encompasses the set individuals of interest. In environmental studies, populations are usually bounded in space and time; e.g., the population of smallmouth bass in Leech Lake, Minnesota on July 1, 2000.

population parameter - a constant term(s) in a mathematic expression, such as a probability density function, that specifies the distribution of individual values in the population. Parameters typically control the location of the center of the distribution (location parameters), the spread of the distribution (scale or dispersion parameters) and various aspects of the shape (shape parameters) of the distribution (see also: **probability density function**).

power of a statistical test -the probability of rejecting the null hypothesis when it is false. Notice that we would like always to reject a false hypothesis; thus, statistical tests with high power (i.e., power >0.80) are desirable. Generally the power of a test increases with the number of individuals in the sample from which the test was computed.

precision - a term applied to the uncertainty in the estimate of a parameter. Measures of the precision of an estimate include its standard error and the confidence interval. Decreasing the value of either leads to increased precision of the estimator.

probability-based sample - a sample selected in such a manner that the probability of being included in the sample is known for every unit on the sampling frame. Strictly speaking, formal statistical inference is valid only for data that were collected in a probability sample.

probability density function (PDF)- for a continuous variable, a curve described by a mathematical formula which specifies, by way of areas under the curve, the probability that a variable falls within a particular range of values. For example, the normal probability density function of the continuous random variable X, is:

$$\frac{1}{s\sqrt{2p}}\exp\left[-\left(\frac{1}{2s^2}\right)(x-m)^2\right]$$

The normal probability density function has two **parameters**, the mean and variance , μ and σ^2. The mean is the location parameter and the variance is the scale parameter; the normal distribution does not have any shape parameters. The graph of the normal probability density function is the familiar "bell curve".

rank - the relative position of a sample value within a sample.

relative frequency - the frequency of occurrence of a given type of individual or member of a group, expressed as a proportion of the total number of individuals in the population or sample that contains the groups. For example, the relative frequencies of 14 bass, 6 bluegill, and 10 catfish in a sample of 30 fish are, respectively: 46.7%, 20.0% and 33.3%.

representative sample - A sample which captures the essence of the population from which it was drawn; one which is typical with respect to the characteristics of interest, regardless of the manner in which it was chosen. While representativeness in this sense cannot be completely assured, probability-based samples are more likely to be representative than are judgement or convenience samples. This is true because only in probability sampling will every population element have a known probability of selection.

sample - a set of units or elements selected from a larger population, typically to be used for making inferences regarding that population.

sampling design - a protocol for the collection of samples from a population, wherein the number, type, location (spatial or temporal) and manner of selection of the units to be measured is specified.

sampling distribution - the expected probability distribution of the values of a statistic that have been calculated from a large number of random samples. For example, the sampling distribution of the ratios of each of the means from 100 samples (each with n=30) to their respective variances will be a t-distribution with 29 degrees of freedom.

sampling error - the difference between a sample estimate and the true population parameter due to random variability in the composition of the sample vs. that of the target population.

sampling frame - the list from which a sample of units or elements is selected.
sampling unit - the members of a population that may be selected for sampling.

significance level (α) - the level of probability at which it is agreed that the null hypothesis will be rejected; α is also the probability of a Type I error.

skewness - a measure of the asymmetry in a distribution, relative to its mean. A right-skewed distribution is composed mostly of small values lying close to the mean but possesses a few values that are much larger than the mean. Conversely, a left-skewed distribution is composed mostly of values lying close to the mean but possesses a few values that are much smaller than the mean.

square-root transformation - a transformation on a variable, X, obtained as, $Y= \sqrt{X}$ or $Y= \sqrt{X + 1/2}$. This transformation is useful for normalizing a discrete variable with a Poisson distribution and/or stabilizing the variance of a variable whose variance is proportional to its mean, prior to statistical analysis with tests (e.g., t-tests) that assume normality and/or variance homogeneity.

standard deviation - the square root of the **variance**.

standard error - the standard error of a sample statistic, θ, (say a sample mean or proportion) is the standard deviation of the values of that statistic computed from repeated sampling of the target population, using the same sampling design (e.g., stratified simple random sampling) and the same sample size, *n*. For example, the standard error of the mean is the sample **standard deviation**/n.

standard normal distribution - a normal distribution whose mean is 0 and whose variance is 1.

statistic - a quantity calculated from the values in a sample (e.g., the sample mean or sample variance)**.**

statistical distribution - a probability distribution used to describe a statistic, a set of observations or a population.

statistical test of hypotheses - a statistical procedure for determining if a sample provides sufficient evidence to reject one statement regarding the population of interest (the null hypothesis) in favor of an alternative statement (the **alternative hypothesis**).

target population - the set of all units or elements, bounded in space and time, about which a sample is intended to produce inferences.

two-sample tests - Statistical tests that evaluate the null hypothesis that there is no difference between a sample statistic (e.g., mean or proportion) in a treatment group or at an ambient monitoring site and the sample statistic in a control group or at a reference site.

Type I error (α) - the error that occurs when a decision maker rejects the null hypothesis when it is actually true. Also called the false rejection decision error, or false positive decision error.

Type II error (β) - the error that occurs when a decision maker accepts a null hypothesis when it is actually false. This is also called the false acceptance decision error, or false negative decision error. The power of a statistical test is 1-β.

variance (population) - the variance of a finite population of N values - $x_1, x_2, ...x_N$ - is simply the average of the squared difference between the individual observations and the population mean.

variance (sample) - the variance of n sample observations is simply the average of the squared differences between the individual observations and the sample mean, divided by (n-1).

variogram - a plot of the sample values of the variance of a spatially referenced variable vs. the corresponding lag distances

APPENDIX D (DRAFT)

INTERVAL ESTIMATORS AND HYPOTHESIS TESTS FOR DATA QUALITY ASSESSMENTS IN WATER QUALITY ATTAINMENT STUDIES

Michael Riggs, Dept. Statistical Research, Research Triangle Institute

Elvessa Aragon, Dept. Statistical Research, Research Triangle Institute

Acknowledgements:

We are grateful to Andy Clayton (RTI), Florence Fulk (EPA/NERL), Laura Gabanski (EPA/OWOW), Susan Holdsworth (EPA/OWOW), Forrest John (EPA/Region 6), and Roy Whitmore (RTI) for providing thoughtful reviews of earlier drafts of this appendix and for suggesting a number of changes which greatly improved the final version.

Appendix D Table of Contents

Appendix D

Interval Estimators and Hypothesis Tests for Water Quality Attainment Studies

D.0 Introduction

This appendix provides formulae for the calculation of inferential statistics and sample size estimates necessary to complete the DQO and DQA processes for design and analysis of water quality attainment studies. These include confidence interval estimators for means, proportions, variances and percentiles, formulae for estimating minimum sample sizes necessary to compute confidence intervals of prespecified half-widths, and instructions for computing one-sample test statistics that can be used to test hypotheses about the attainment of water quality standards. Examples of the computation of all estimators and test statistics accompany the text. Special attention is paid to the exact binomial and exact hypergeometric tests.

Although all of the computations described in this appendix can be completed with statistical tables that are readily available in standard introductory texts (e.g., Lindley and Scott 1995; Steel et al. 1996) or on the Internet (e.g., http://www.epa.gov/quality1/qa_docs.html) and a hand calculator, in practice, they are usually done with the aid of statistical software. EPA provides software (DEFT) that may be used in the DQO process for sample size estimation for 1-sample t-tests and z-tests. EPA provides an additional software package (DataQuest) that can be used to carry out these tests and the large-sample Wilcoxon signed ranks test. Both packages can be downloaded from, http://www.epa.gov/quality1/qa_docs.html.

Although EPA does not endorse specific commercially available software, the following statistical packages are widely available at many academic and public research institutions. Perhaps the most comprehensive and widely used are SAS (http://www.sas.com/), SPLUS (http://www.insightful.com/products/default.asp), and its freeware counterpart R (http://www.r-project.org/). All of these packages provide well documented procedures and functions for a wide variety of statistical tests and estimators including those described in this appendix; however, they all require users to be conversant in their respective programming languages. By contrast, the SPLUS EnvironmentalStats package (http://www.probstatinfo.com/; Millard and Neerchal 2001), specifically designed to support the DQO and DQA processes, has an easy to use GUI interface through which all of the estimators, tests and sample size computations in this appendix may be produced with a few mouse clicks. Similarly, PASS (http://www.ncss.com/pass.html) and CIA (http://www.som.soton.ac.uk/cia/; Altman et al. 2000) are relatively inexpensive, menu-driven software for (respectively) power/sample size estimation and confidence interval estimation.

D.1. The Confidence Interval as a Tool for Specifying/Controlling Decision Error Rates

Formulae for constructing two-sided $100\times(1-\alpha)\%$ confidence intervals from sample estimates of population means, binomial proportions, and variances are presented in Box 1a and are illustrated with sample calculations in Box 1b of this Appendix. These and all other confidence interval formulae in this section are extensions of the general formulae for 1- and 2-sided confidence intervals that were presented in Box 2 of Appendix C. General characteristics, utility

and interpretation of confidence interval estimates are reviewed in Section C.1.5 of Appendix C. All of the 2-sided computations illustrated in this section can be easily extended to the case of 1-sided estimators on the basis of the relationships defined in Box 2 of Appendix C.

The expressions for confidence intervals on means and proportions (Box 1 of this Appendix) can be rearranged to solve for the sample size (n). Sample size formulae permit one to estimate *a priori* (i.e., during the DQO process) the minimum number of sampling units that must be collected in order to yield a confidence limit whose width is less than a specified maximum. For example the sample size required for $100\times(1-\alpha)\%$ two-sided confidence interval on the mean, with a half-width of W, can be obtained from the iterative solution of the first equation in Box 2a. Similarly, the sample size required for a $1-\alpha\%$ confidence interval on a binomial proportion with a half-width of W can be calculated with the second expression in Box 2a. Sample sizes for one-sided confidence intervals can be obtained by replacing the $1-\alpha/2$ t- and z-statistics in Box 2a with the corresponding $1-\alpha$ statistics. In either case, sample size calculations based on these formulae require the investigator to specify her desired maximum allowable Type I error rate, the standard error on the estimate and her desired maximum half width for the confidence interval. Optimally, the standard error should be obtained from either a pilot study or some previous study on the same target population (e.g., a previous assessment); failing that, the investigator may choose to use a standard error from a published study of a similar population. When α and SE are fixed, specifying the maximum acceptable confidence interval half-width is comparable to controlling the Type II error rate in a statistical hypothesis test.

Consider the case of an investigator who desired 95% confidence limits on the acute CMC of selenite from a sample of n 1-liter volumes taken from a stream reach, such that the confidence interval half-width (W) would be ≤ 20 µg/L. A pilot study indicated that the standard deviation for selenite concentration in the stream was 200. Applying the first equation in Box 2a, he determined that he needed to collect 387 1-liter volumes from the stream reach to meet this requirement. Alarmed, he recomputed n, this time specifying w=50 µg/L and was relieved find the new sample size requirement of 64 volumes. This demonstrates the point made earlier; simultaneous control of α and β to low levels requires extremely large sample sizes. The only way he could maintain an α of 0.05, without increasing the sample size, was to increase the confidence interval width, thereby increasing the Type II error rate and decreasing the precision of his estimate. However, if he is willing to accept a higher Type I error rate, say α=0.15, he could maintain a half-width of $\leq$ 20 µg/L on an 85% confidence interval with a sample size of only n=209 1-liter volumes, as compared to n=387 for a 95% confidence interval with the same half-width. Detailed discussion of the complex interrelationships among specified -levels and half-widths, variances and other factors that affect confidence intervals and hypothesis tests are taken up in Sections C.3.1 and C.3.2 of Appendix C.

Box 1-a : Computation of 100 x (1-a)% Confidence Intervals for Sample Estimates

Population Mean

Suppose that $c_1, c_2 \ldots, c_n$ represents a random sample of $n \geq 30$ data points from an essentially infinite population of sampling units.

Step 1. Calculate the sample mean $\bar{c}$ and the sample variance s^2 (see Appendix C, Box 1-a and 1-b).

Step 2. Use the table of percentage points of student's t-distribution to determine the value $t_{1-a/2,n-1}$ such that 100 x (1-a/2)% of the student t-distribution with n-1 degrees of freedom is below $t_{1-a/2,n-1}$.

Step 3. The 100 x (1-a)% confidence interval for the population mean is bounded by the following upper and lower limits:

$$\bar{c} - t_{1-a/2,\,n-1}\sqrt{\frac{s^2}{n}} \text{ to } \bar{x} + t_{1-a/2,\,n-1}\sqrt{\frac{s^2}{n}}\,.$$

Step 4. If n is ≥ 60, the value of $t_{1-a/2,\,n-1}$ may be replaced by the value $z_{1-a/2}$ of the standard normal distribution (the table of percentage points of student's t-distribution).

Population Binomial Proportion

Suppose that c is a dichotomous random variable with values 0, 1 (denoting, for example, the presence of absence of some observable characteristic). Let $c_1, c_2 \ldots, c_n$ represent a random sample of n data points from a population of c values. If n is ≥ 50, an approximate $100 \times (1 - a/2)\%$ confidence interval can be computed as follows:

Step 1. Calculate the sample proportion p: $p = \frac{1}{n}\sum_{i=1}^{n} c_i$

Step 2. Use the table of the cumulative probabilities of the standard normal distribution to determine the value of $z_{1-a/2}$ such that $100 \times (1 - a/2)\%$ of the standard normal distribution is below $z_{1-a/2}$.

Step 3. The large-sample $100 \times (1 - a/2)\%$ confidence interval for the population proportion is bounded by the following lower and upper limits:

$$p - z_{1-a/2}\sqrt{\frac{p(1-p)}{n}} \text{ to } p + z_{1-a/2}\sqrt{\frac{p(1-p)}{n}}\,.$$

Population Variance

Suppose that $c_1, c_2 \ldots, c_n$ represents a random sample of n data points from a population of normally distributed values for c.

Step 1. Calculate the sample variance s^2 (see Appendix C, Box 1-a).

Step 2. Use Table of critical values of the chi-square distribution to determine the value of $c^2_{1-a/2,\,n-1}$ such that) $100 \times (1 - a/2)\%$ of the chi-square distribution with n-1 degrees of freedom below $c^2_{1-a/2,\,n-1}$. Similarly, determine the value $c^2_{a/2,\,n-1}$ such $100 \times (a/2)\%$ of the chi-square distribution with n-1 degrees of freedom is below $c^2_{a/2,\,n-1}$.

Box 1-a : (Continued) Calculating of 100 x (1-a)% Confidence Intervals

Step 3. The 100 x (1-a)% confidence interval for the population mean is bounded by the following upper and lower limits:

$$\frac{(n-1)s^2}{c^2_{1-a/2,\,n-1}} \quad , \quad \frac{(n-1)s^2}{c^2_{a/2,\,n-1}} .$$

Note :
The **width** of the confidence interval is the distance between the upper and lower confidence limits. The confidence intervals for the mean and proportion are symmetric; that is, the distance between the estimates and their upper and lower bounds (**half-width**) are equal. The confidence interval for the variance is asymmetric.

Box 1-b : Computation of 100 x (1-a)% Confidence Intervals for Sample Estimates

Consider the random sample of 10 turbidity measurements from Appendix C, Box 1-b. Suppose that the population of turbidity measurements is normally distributed. (Later sections will discuss how to determine if a normal distribution may be assumed for a population.) To calculate 95% confidence intervals for the mean, proportion and variance, follow the steps outlined in Box 1-a of this Appendix.

Population Mean

Step 1. The sample mean and sample variance were calculated in Box 1-b:

$$c = 101.3 \quad \text{and} \quad s^2 = 8111.7889$$

Step 2. Since $df = 10 - 1 = 9$, and $1\text{-}a = 0.95$, $a = 0.05$, $1 - a/2 = 1 - 0.05/2 = .0975$, use the Table of percentage points of student's t-distribution to determine the value $t_{0.975,9}$ of the distribution with 9df. This value is 2.262.

Step 3. Calculate the lower and upper limits.

$$101.3 - 2.262\sqrt{\frac{8111.7889}{10}} = 101.3 - 64.42 = 36.88$$

$$101.3 + 2.262\sqrt{\frac{8111.7889}{10}} = 101.3 + 64.42 = 165.72$$

The 95% confidence interval for the population mean is (36.88, 165.72). Note that the width of the confidence interval is 64.42.

Population Proportion

Note that the sample size *n=10* is not large enough to use the steps outlined in Box 1-a of this Appendix. The exact binomial confidence interval discussed in a later section will appropriate for small sample sizes.

For purposes of showing a sample calculation, suppose that we have a sample of 75 turbidity measurements from the same population.

Step 1. Suppose that the sample proportion was calculated to be *p=0.18*.

Step 2. Use the Table of the cumulative probabilities of the standard normal distribution to determine the value $z_{0.975}$ of the standard normal distribution. This value is 1.96.

Step 3. The lower and upper limits are

$$0.18 - 1.96\sqrt{\frac{0.18(1-0.18)}{10}} = .018 - 0.24 = -0.06$$

$$0.18 + 1.96\sqrt{\frac{0.18(1-0.18)}{10}} = .018 + 0.24 = 0.42$$

Note that the lower limit is -0.06. Since proportions cannot be negative, the lower limit becomes 0. The 95% confidence interval for the population proportion is (0, 0.42). Note that the width of the confidence interval is 0.24.

Box 1-b: (Continued) Examples for Calculating of 100 x (1-a)% Confidence Intervals

Population Variance

Step 1. From Box 1-b of Appendix C, s^2=8111.8.

Step 2. Use the Table of critical values of the chi-square distribution to find the values $c^2_{0.975,9}$ and $c^2_{0.025,9}$ of the chi-squared distribution with 9 df. These values are 19.023 and 2.700, respectively.

Step 3. The lower and upper 95% confidence limits on the sample estimate of the variance (s^2) of the turbidity values in Mermentau River, June 1980 - April 2000, are

$$\frac{(10-1)8111.8}{19.02} = 3837.8 \quad \text{and} \quad \frac{(10-1)8111.8}{2.70} = 27039.3$$

The 95% confidence interval for the population variance is (3837.7806, 27039.2963). Note that the confidence interval is asymmetric; the lower confidence interval has a width of 4274 (8111.8-3837.8), while the width of the upper confidence interval is 18,927.5 (27039-8111.8). This reflects the asymmetry of the chi-square distribution which is used to model the variance. Note also the extreme width of the overall confidence interval (23,201.5). This demonstrates a general attribute of the population variance: unless s^2 is small, it is very difficult to obtain good estimates of s^2 from small samples. The population variances of most ambient environmental parameters (e.g. turbidity) and natural population parameters will usually be quite large.

Box 2-a : Minimum Sample Size Requirements for Estimating a Population Mean of Binomial Proportion with 100 x (1-a)% Confidence.

Population Mean

Step 1a. Decide on the level of confidence (1-a) and maximum acceptable half-width (W) of the 100 x (1-a)% confidence that you will be calculating.

Step 2a. Iteratively search for the smallest value of n that satisfies the equality:

$$n = \left(t_{1-a/2,\, n-1} c \, \frac{s}{W} \right)^2 \qquad \text{where } s = \sqrt{s^2}$$

Step 3a. Typically, *n* will not be an integer. Round up the calculated value of *n* to the next integer to obtain the minimum sample size required to insure a 100 x (1-a)% confidence interval whose half-width is no larger than W.

Population Proportion

Step 1b. Decide on the level of confidence (1-a) and maximum acceptable half-width (W) of the 100 x (1-a)% confidence interval you will be calculating.

Step 2b. Solve the following algebraic expression for *n*:

$$n = \frac{1}{\left(\sqrt{z^2_{1-a/2} p(1-p) + 2W} - z_{1-a/2} \sqrt{p(1-p)} \right)^2}$$

Step 3b. Round up the calculated value of *n* to the next integer to obtain the minimum sample size required to insure a 100 x (1-a)% confidence interval whose half-width is no larger than W.

Note:
While the minimum sample size "n" appears on both sides of the equation used to get *n* for the mean (2a of this Appendix), it appears only on the left side of the equation that is used to get a minimum *n* for proportion (2b of this Appendix). The practical consequence of this is that we can use basic algebra to solve for the *n* needed to construct a confidence interval for a proportion, but we must use a trial-and-error approach (i.e. iteration) to get the corresponding minimum *n* for the case of the mean. Ideally, one should write a computer program with a "do-loop" to generate a large number of candidate solutions for *n*, computed over a wide range of df values (i.e., n-1) for the t-statistic on the left hand side of the equation in step 2a. However, for those without access to a computer, we provide in Box 2-b of this Appendix, a more tedious solution that can be done with the aid of a hand calculator.

The confidence interval formulae in Box 1a require the assumption that repeated sample estimates of the population parameters be normally distributed about the true population parameter. This assumption has been proven to hold for a large number of sample estimators that are based on an estimation procedure called **maximum likelihood**. In general, the normality assumption will hold for the sample means and proportions based on n > 20-30. When sample sizes are <20, different estimation methods should be considered. For example, confidence intervals for binomial proportions from samples with sample sizes (n) such that $n \times p = 5.0$ and/or $n \times (1-p) = 5.0$, should be computed using exact binomial methods. Exact confidence intervals and associated minimum sample size requirements can be obtained from tables (e.g., CRC Basic Statistical Tables) or from widely available statistical software packages such as PASS, SPLUS EnvironmentalStats, and SAS (O'Brien 1998; http://www.bio.ri.ccf.org/power.html).

Two-sided confidence intervals for the sample median or for sample percentiles (e.g., the 95th percentile) can be computed using nonparametric methods; so-called because they don't assume a particular underlying parametric distribution such as the normal. Like the above-described methods, the nonparametric methods have both large-sample (n>20) and small sample exact forms. Nonparametric exact confidence intervals and minimum sample size requirements can be obtained from tables in nonparametric statistics books (e.g., Hollander and Wolfe 1999) or from statistical software packages [e.g., SAS (PROC FREQ and PROC NPAR1WAY)]. The large sample formulae for 1-α% confidence intervals on the population median or on any of its percentiles are based on the ranks of the data in the sample. The ranks are obtained by first sorting the data from smallest to largest, finding the value which corresponds to the median or desired percentile and then computing the ranks of the observations which correspond to the lower (r) and upper (s) 1-α% bounds on the sample estimate. The formulae for computing the median and other population percentiles and r and s for their confidence intervals are shown in Boxes 3 and 4, respectively.

For example, consider the 244 monthly turbidity values in Table 3 of Appendix C. The data are sorted from smallest to largest beginning with the first value in row one and increasing from left to right within the row. The largest values are in the last row (row 49). Because the sample size (244) is an even number, the median is computed as the mean of the middle two observations (i.e., 75 and 76, bold in row 24). To compute the 2-sided 90% confidence interval on the median, we apply the first two equations in Box 4. First, we find Z associated with 1-(0.10/2), which is 1.645. Next, we solve the equations to find that the ranks of the 90 % lower and upper bounds on the median turbidity correspond to 109.15 and 135.8. However, ranks must be integers, so r needs to be rounded *down* to the nearest integer (109) and s needs to be rounded *up* to the nearest integer (136). The turbidity values corresponding to these ranks (bold in Appendix C Table 3) are 70 and 80; thus:

2-sided 90% confidence interval on the median = 75.5 (70.0, 80.)

In addition, using the equations in Box 5a, we can easily write the 1-sided 90% confidence intervals for the median turbidity:

Box 3-a : Sample Percentiles and the Sample Median

In a population or sample, a percentile is the data value that is greater than or equal to a given percentage of all the data values. For instance, the 90th percentile is the data value that is greater than or equal to 90% of all the data values. The 50th percentile is more commonly called the sample median.

Let $c_1, c_2 \ldots, c_n$ represent a random sample of *n* population units, ordered from the smallest to the largest values of c_i. The pth percentile, is the value of c associated with the $c(p)^{th}$ ordered value from the sample size *n*; i.e. $c(p)$ is the rank of c value which is $\geq$ p-percent of the other values in the sample. $c(p)$ can be determined by the following simple 4-step process:

Step 1. Order the data values from smallest to highest and label these ordered values $c_{(1)}, c_{(2)} \ldots, c_{(n)}$. Thus $c_{(1)}$ is the smallest value, $c_{(2)}$ is the second smallest, and $c_{(n)}$ is the largest.

Step 2. Calculate np/100. Separate the integer part of the result. That is,

$$\frac{np}{100} = i + f, \qquad \text{where } i \text{ is the integer part and } j \text{ is the fraction part.}$$

For instance, 135.78=135+0.78.

Step 3. If *f=0*, then $c(p) = \frac{c_{(i)} + c_{(i+1)}}{2}$. Otherwise, $c(p) = c_{(i+1)}$.

Step 4. For the sample median, *p=50.* Hence, *n*(50)/100=*n*/2; i.e. the value of c associated with $n/2^{th}$ ordered sample value is the sample median.

Box 3-b : Example for Calculating Sample Percentiles and the Sample Median

Consider the 10 turbidity measurements from Box 1-b in Appendix C : 34, 58, 87, 95145, 14, 38, 62, 95, 160, 320.

Step 1. The ordered values are: 14, 34, 38, 58,62, 87, 95, 145, 160, 320.
Step 2. Calculate the 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles.

$$p = 25 : \frac{(10)(25)}{100} = 2.5 = 2 + 0.5$$
$$p = 50 : \frac{(10)(50)}{100} = 5 = 5 + 0$$
$$p = 75 : \frac{(10)(75)}{100} = 7.5 = 7 + 0.5$$
$$p = 90 : \frac{(10)(90)}{100} = 9 = 9 + 0$$

Step 3. Determine y(p).

p	i	f	$y(p)$
25	2	0.5	$\frac{c_{(2)} + c_{(3)}}{2} = \frac{(14+34)}{2} = 2$
50	5	0	$c_{(5)} = 62$
75	7	0.5	$\frac{c_{(7)} + c_{(8)}}{2} = 95 + 145/2 = 120$
90	9	0	$c_{(10)} = 320$

Box 4-a : Large-Sample 100 x (1-a)% Confidence Intervals for Percentiles

Population Median

Let $c_{(1)}, c_{(2)}..., c_{(n)}$ represent the ordered values in a random sample of *n* units, such that ≤ 20 from a population. Calculate

$$r = \frac{n}{2} - \left(z_{1-a/2} c \frac{\sqrt{n}}{2} \right)$$

$$s = 1 + \frac{n}{2} + \left(z_{1-a/2} c \frac{\sqrt{n}}{2} \right)$$

Round *r* down to the next lower integer and round *s* up to the next highest integer. Find the r^{th} and s^{th} ordered values in the sample. The large-sample 100 x (1-a)% confidence interval for the median is $\left(c_{(r)}, c_{(s)} \right)$.

p^{th} Population Percentile

Let $c_{(1)}, c_{(2)}..., c_{(n)}$ represent the ordered values in a random sample of *n* units from a population. Calculate

$$r = \frac{np}{100} - \left(z_{1-a/2} c \frac{\sqrt{np(100-p)}}{100} \right)$$

$$s = 1 + \frac{np}{100} + \left(z_{1-a/2} c \frac{\sqrt{np(100-p)}}{100} \right)$$

Round *r* down to the next lower integer and round *s* up to the next higher integer. Find the r^{th} and s^{th} ordered values in the sample. The large-sample 100 x (1-a)% confidence interval for the p^{th} percentile is $\left(c_{(r)}, c_{(s)} \right)$.

Box 4-b : Examples for Calculating Large-Sample 100 x (1-a)% Confidence Intervals for Percentiles

Consider a random sample of 25 turbidity measurements from the Louisiana River between June 1980 and April 2000. The ordered (upper left to lower right) sample values are:

14	26	34	38	40
45	50	50	58	62
70	75	76	80	87
95	115	125	130	145
160	175	210	245	320

Calculate 95% confidence intervals for the median and the 90th percentile.

From the table of the cumulative probabilities of the standard normal distribution, $z_{0.975} = 1.96$.

For the median:
Using the steps in Box 3-a, the sample median is determined to be the 13th ordered value, which is 76.

$$r = \frac{25}{2} - \left(1.96c\frac{\sqrt{25}}{2}\right) = 12.5 - 4.9 = 7.6 \quad \text{rounded down to } 7$$

$$s = 1 + \frac{25}{2} + \left(1.96c\frac{\sqrt{25}}{2}\right) = 13.5 + 4.9 = 18.4 \quad \text{rounded up to } 19$$

The 7th ordered value is 50, and the 19th ordered value is 130. Hence, the 95% confidence interval for the median is (50, 130).

For the 90th percentile:

Using the steps in Box 3-a, the sample 90th percentile is determined to be $\left(c_{(22)} + c_{(23)}\right)/2 = (175 + 210)/2 = 192.5$.

$$r = \frac{25}{2} - \left(1.96c\frac{\sqrt{25}}{2}\right) = 12.5 - 4.9 = 7.6 \quad \text{rounded down to } 7$$

$$s = 1 + \frac{25}{2} + \left(1.96c\frac{\sqrt{25}}{2}\right) = 13.5 + 4.9 = 18.4 \quad \text{rounded up to } 19$$

The 19th ordered value is 130. Since n=25, there is no 27th ordered value, so the largest ordered value, 320, becomes the upper limit. Hence, the 95% confidence interval for the 90th percentile is (130, 320).

Lower 1-sided 90% confidence interval on the median = 75.5 (70.0, $+\infty$)
Upper 1-sided 90% confidence interval on the median = 75.5 ($-\infty$, 80.0)

Because the median and the percentiles are based on ranks, the minimum sample size requirements for their estimation needs to be stated a little differently than for means, proportions or variances. Applying results from the theory of order statistics, we can calculate the minimum number of sampling units that is required for a sample to contain at least one value $\geq$ the value of the population median (second equation in Box 5a) or, if we desire, of the population p^{th} percentile (first equation in Box 5a), with 1-α% confidence. For example, if an investigator desires to have 95% confidence that a sample of sampling units from a lagoon contains at least one sampling unit whose selenium concentration is as large or larger than the 95^{th} percentile of selenium concentration in the universe of all possible sampling units from the lagoon, the first equation in Box 6a dictates that his sample must contain at least 59 sampling units.

A final point to consider when using nonparametric methods is that the investigator does not have as much control over the Type II error rates as he does with the parametric methods. This is because he cannot specify a minimum confidence interval half-width; all he can control is the α-level, which he does by specifying a 1-α% confidence level. However, it is possible to obtain minimum sample size estimates for the median, which control for both α and β by applying methods based on the Wilcoxon Signed-Rank Test (both exact and large-sample versions; Hollander and Wolfe 1999).

The methods described to this point allow the specification of acceptable decision error rates based on specification of the 1-α% confidence level and, conditional on such specification, fixing the half-width of the confidence interval (and thereby the Type II error rate) by specifying a minimum sample size. For a given 1-α% confidence level and a standard value C, against which the sample estimate is to be compared, specifying (for example) a maximum allowable half-width of 0.10 x C will provide substantially more statistical power (i.e., 1-β) than will a specified half-width of 0.50 x C.

D.2 Parametric one-sample tests on means

The use of the one-sample t-test to evaluate attainment vs. impaired hypotheses involving the mean of a continuous water quality variable (e.g., turbidity) against a criterion value (e.g., the maximum allowable turbidity value in a stream) was discussed in detail in Sections C.2.1 Appendix C. The assumptions required by the test were listed in section C.3.1 and graphical methods for their verification were illustrated in Section C.2.1. In this section we briefly review the features and assumptions of the t-test and then focus on the problem of employing the test for analysis of data that are highly autocorrelated and confounded by significant seasonal effects, two commonly encountered problems in water quality studies.

The general form of the t-statistic is provided in Box 6. When the null hypothesis is true and the sampling units have been independently sampled from a population in which the attribute values are normally distributed, t will follow a t-distribution with n-1 degrees of freedom (df). But if the alternative hypothesis is true, it will follow a noncentral t-distribution with df=n-1.

Box 5-a : Minimum Sample Size Requirements for Constructing 100 x (1-a)% Confidence Intervals for Percentiles

If we want to have to 100 x (1-a)% confidence that a random sample of size *n* from a target population contains a one value of c (e.g., turbidity) that is at least as large as the pth population percentile of c , the minimum sample size can be computed as:

$$n = \frac{\ln\left[1-(1-a)\right]}{\ln\left(\frac{p}{100}\right)}$$

For the median (p=50), this becomes

$$n = \frac{\ln\left[1-(1-a)\right]}{$$

Box 5-b : Examples for Calculating the Minimum Sample Size Requirements for Constructing 100 x (1-a)% Confidence Intervals for Percentiles

Suppose we are in a monitoring situation wherein we need to be 95% certain that we regularly obtain samples from the sediment that contain at least 1 sediment aliquot (i.e., sampling unit) whose concentration of selenium is at or above the median of selenium concentrations in the reservoir that is being monitored? At or above the 90th percentile?

Question 1: How many sample units (i.e., aliquots of sediment) must the sample contain in order to be 95% confident that the sample contains at least one aliquot whose selenium concentration is as large as, or larger than, the median (=50th percentile) of the selenium concentration of the entire target population?

Letting p=0.50 and $a = 0.01$ in the equation in Box 3-a, *n* is computed as:

$$n = \frac{\ln(1-0.95)}{\ln(0.50)} = \frac{\ln(0.05)}{\ln(0.50)} = 4.32 \qquad \text{rounded up to 5}$$

Question 2: Similarly, how many sampling units must the sample contain in order to be 95% confident that the sample contains at least one aliquot whose selenium concentration is as large as, or larger than, the 90th percentile of the selenium concentration of the entire target population?

Letting p=0.90 and $a = 0.01$ in the equation in Box 3-a, *n* is computed as:

$$n = \frac{\ln(1-0.95)}{\ln(0.90)} = \frac{\ln(0.05)}{\ln(0.90)} = 28.43 \qquad \text{rounded up to 29}$$

Box 6-a: One-Sample t-Test for the Mean (One-sided Case)

Assumptions

1. The distribution of the attribute X is normal with population mean μ and variance s^2.
2. $X_1, X_2, \ldots, X_n$ is an independent sample of n individuals from the target population.

Let μ_0 be the fixed criterion against which the population mean μ is compared. Consider each of the hypothesis pairs:

Case 1: $H_0 : \mu = \mu_0$ vs. $H_a : \mu > \mu_0$
Case 2: $H_0 : \mu = \mu_0$ vs. $H_a : \mu < \mu_0$

For either case, when the two assumptions hold, a one-sample t-test for the null hypothesis vs. the alternative hypothesis can be performed as follows:

Step 1. Select the significance level a. (Typical values of a are 0.05 and 0.01.)
Step 2. Calculate the sample mean $\bar{x}$ and the sample variance s^2 (see Box 1-a).
Step 3. Use the table of percentage points of student's t-distribution to find the value $t_{1\text{-a, n-1}}$ such that (1-a)x100% of the Student t distribution with n-1 degrees of freedom is below $t_{1\text{-a, n-1}}$.
Step 4. Calculate the test statistic:

$$t_c = \frac{\bar{x} - m_0}{\sqrt{\frac{s^2}{n}}}$$

where $\bar{x}$ = the sample mean of the measured attribute, X
s^2 = the sample variance of the measured attribute, X
n = the number of sampling units in the sample
μ_0 = the fixed criterion against which the population mean is compared.

Step 5. Compare t_c with $t_{1\text{-a, n-1}}$.
Case 1: If $t_c > t_{1\text{-a, n-1}}$ then reject H_0.
If $t_c = t_{1\text{-a, n-1}}$ then reject H_0.
Case 2: If $t_c < t_{1\text{-a, n-1}}$ then reject H_0.
If $t_c = t_{1\text{-a, n-1}}$ then reject H_0.

Box 6-b: Example for Performing a One-Sample t-Test for the Mean

Consider a random sample of 35 turbidity measurements from the Mermentau River between June 1980 and April 2000. The sample values are:

34	58	80	87	145	245	26
40	50	75	130	175	14	38
62	76	95	160	45	115	210
320	50	70	125	8	432	52
26	19	32	80	85	170	352

It is desired to test the null hypothesis that the mean monthly turbidity measurement is no more than 150 NTU vs. the alternative the mean is greater than 150 NTU.

$$H_0 : \mu \leq 100 \quad \text{vs.} \quad H_a : \mu > 100$$

Step 1. The desired significance level is a=0.05.

Step 2. Using the equations in Box 1-b of Appendix C, the sample mean and sample variance are calculated to be $\bar{x}$ = 108.1 and s^2 = 9924.70.

Step 3. With a=0.05, 1-a/2=1-0.025=0.975 and df=35-1=34. The table of percentage points of student's t-distribution yields $t_{0.975,\,34}$=2.032.

Step 4. Calculate the test statistic.

$$t = \frac{108.1 - 150}{\sqrt{\frac{9924.70}{35}}} = \frac{-41.9}{16.84} = -2.49$$

Step 5 Since -2.49 < 2.032, the null hypothesis cannot be rejected. Accept the null hypothesis and conclude that the true mean monthly turbidity seems to be less than or equal to 150 NTU.

While many things in nature tend to be normally distributed (e.g., weights of organisms), environmental data are very commonly log-normally distributed. Thus, *before* computing a t-test, it is important that the normality of the distribution of the sampling units be assessed using the methods described in Sections C.2.1 and C.2.2. If the data are approximately normal they can be analyzed as is; if not, then a suitable transformation (e.g. log-transformation) should be applied to the data before carrying out the t-test. Normality of the transformed variable can be verified using Q-Q plots as illustrated in C.2.2.

Frequently, good sampling and/or experimental design will be sufficient to insure independence among the sampling units. However, as pointed out in Section C.2.5, inherent temporal autocorrelation may be so strong that it will not be possible to design it away without substantial loss of data. We can illustrate the problem of strong autocorrelation and one of its possible solutions with some of the Mermentau River turbidity data. Recall that the turbidity data are approximately lognormal and autocorrelated for lags other than 3, 8 or 9 months. Suppose that we desire to test the null hypothesis that the mean turbidity during the most recent 3-year period (i.e., 1997-1999, inclusive) is less than or equal to the 150 NTU criterion vs. the 1-sided alternative that it is greater than 150 NTU. This requires that we consider the log-transformed values of the 36 monthly turbidity measurements that are listed in the fourth column of Table 1.

The first step is to plot the turbidity time series and the correlogram of the log-turbidity, as was done in the example in Section C.2.5 of Appendix C. These two plots (Figs. 1A and 1B) confirm the existence of both seasonality and significant autocorrelation in the 3-year time series. Not surprisingly, the patterns are quite similar to those of the 20-year series (Appendix C, Fig. 11). Significant autocorrelations occur between measurements that were taken 1, 2, 6, 7, 8, 18 and 19 months apart. The repeating annual pattern indicates strong seasonality in the series. In fact, the assessment of autocorrelation is difficult to make in the face of either seasonality or long-term trends. Thus it is desirable to remove their effects from the series before making a final of interpretation of the correlogram.

Seasonality and long-term trend are defined as fixed effect deviations from the mean of a time-series and can be estimated and removed from the series by the fitting of a two-way **ANOVA model**:

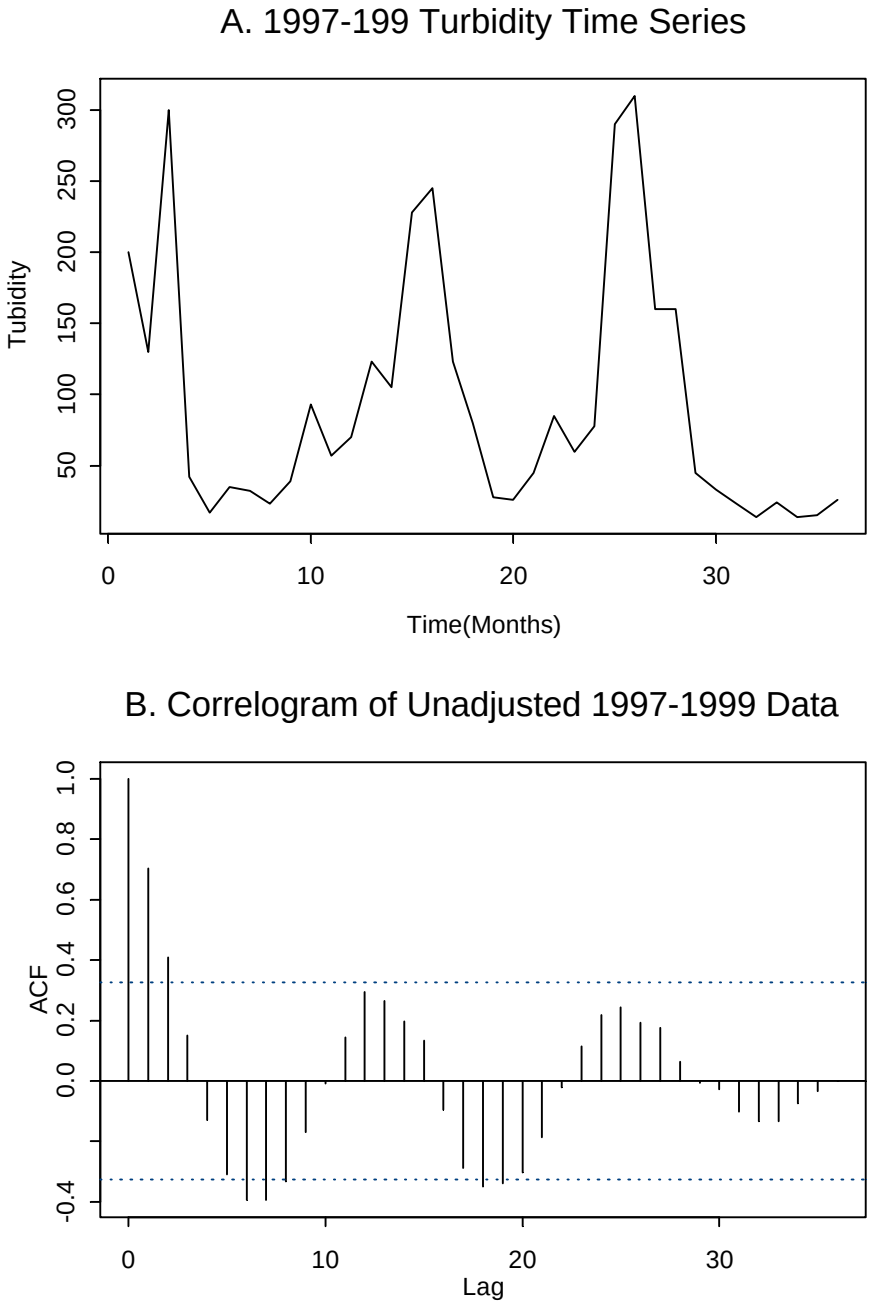
$$y_{ij} = m_0 + b_i Year_i + b_j Month_j + e_{ij} \qquad (1)$$

This model says that each observed monthly log-turbidity value (y_{ij}) is due to the 3-year mean of the time series (μ_0) + an effect due to the particular year in which it was taken (e.g., a dry year or a wet year) + the effect of the month of the year in which it was taken + random error. If there had been increasing development of the lands adjacent to the Mermentau River over the 1997-1999 period this may well have led to increasing erosion and runoff with corresponding annual increases in turbidity. If this had been the case, the year effect would have estimated the 3-year trend. Similarly, month-to-month differences that are consistent across years are an indication of seasonality and are estimated in the ANOVA model by the month effects. A cursory examination of the raw data (Table 1, column 3) suggests the presence of seasonality but not trend. The final term in the model, represented by the Greek epsilon (ε_{ij}), is what is left over after removing trend and seasonality from the overall 3-year average of the log-turbidity measurements. Thus, the ε_{ij} , called residuals, are the seasonally adjusted values that we desire.

Table 1. Intermediate calculations for a seasonally adjusted ttest on 1997-1999 Turbidity data.

YEAR	MONTH	Turbidity (NTU)	Log of Turbidity	Adjusted Log of Turbidity	Squared Adjusted Values	Adjusted x Lag1 Adj. Values
1997	1	200	5.298	0.070	0.005	.
	2	130	4.868	-0.187	0.035	-0.013
	3	300	5.704	0.332	0.110	-0.062
	4	42	3.738	-1.002	1.005	-0.333
	5	17	2.833	-0.953	0.908	0.955
	6	35	3.555	-0.225	0.050	0.214
	7	32	3.466	0.186	0.035	-0.042
	8	23	3.135	0.156	0.024	0.029
	9	39	3.664	0.146	0.021	0.023
	10	93	4.533	0.693	0.480	0.101
	11	57	4.043	0.459	0.211	0.318
	12	70	4.248	0.325	0.106	0.149
1998	1	123	4.812	-0.733	0.537	-0.239
	2	105	4.654	-0.717	0.514	0.526
	3	228	5.429	-0.258	0.067	0.185
	4	245	5.501	0.445	0.198	-0.115
	5	123	4.812	0.710	0.504	0.316
	6	80	4.382	0.286	0.082	0.203
	7	28	3.332	-0.264	0.070	-0.075
	8	26	3.258	-0.038	0.001	0.010
	9	45	3.807	-0.028	0.001	0.001
	10	85	4.443	0.286	0.082	-0.008
	11	60	4.094	0.194	0.038	0.056
	12	78	4.357	0.117	0.014	0.023
1999	1	290	5.670	0.663	0.440	0.078
	2	310	5.737	0.904	0.818	0.600
	3	160	5.075	-0.074	0.005	-0.067
	4	160	5.075	0.557	0.311	-0.041
	5	45	3.807	0.243	0.059	0.135
	6	33	3.497	-0.061	0.004	-0.015
	7	23	3.135	0.078	0.006	-0.005
	8	14	2.639	-0.118	0.014	-0.009
	9	24	3.178	-0.118	0.014	0.014
	10	14	2.639	-0.979	0.958	0.115
	11	15	2.708	-0.653	0.427	0.640
	12	26	3.258	-0.443	0.196	0.289
		Totals	148.400		8.348	3.955

Fig. 1. (A) Time series plots of 36 monthly turbidity measurements from the Mermantau River, January 1997- December 1999. (B) Correlograms showing seasonality and autocorrelation of log-transformed turbidity measurements.



The simplest way to obtain the residuals is to fit the two-way ANOVA model with a statistical software package (e.g., SAS, SPLUS, etc.). These software packages will automatically compute the residuals and save them in a data set, which the analyst can use for the remaining calculations in this section. The values in column 5 of Table 1 (Adjusted Log of Turbidity) are the residuals from the fitted two-way ANOVA model of the log-turbidity data.

Alternatively, if such software is not available, the analyst can easily do the necessary computations with a hand calculator, as follows:

1. Compute the 1997, 1998 and 1997 annual means of the log-turbidity data (4.090, 4.407, 3.686)
2. Subtract from each observed log-turbidity value, its corresponding annual mean
3. Compute the twelve monthly means of the year-adjusted values computed in Step 3 (1.138, 0.0964, 1.281, 0.650, -0.304, -0.311, -0.8101, -1.111, -0.572, -0.250, -0.597, -0.167)
4. Compute the residual for each of the year-adjusted values and the corresponding monthly deviations computed in Step 3 by subtracting the appropriate annual and monthly mean from observed log-turbidity value. The following illustrates the computation of the residual for Month 1 of 1997 is computed as:

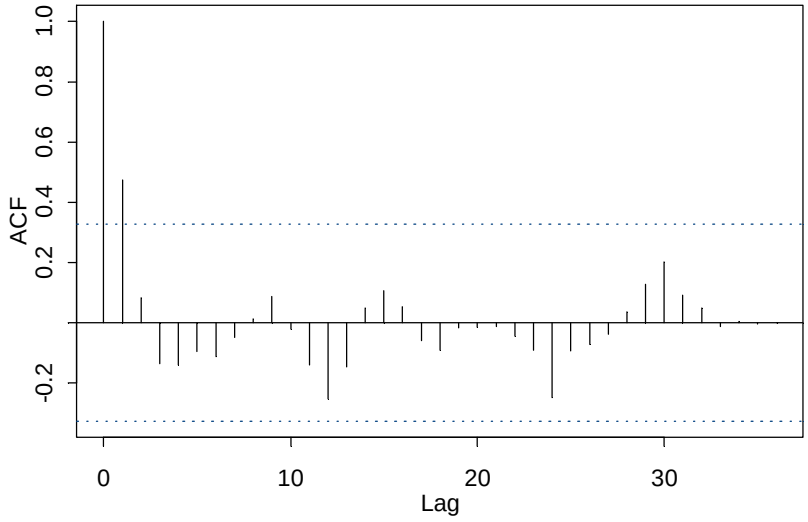
Jan. 1997 log-turbidity value	– 1999 mean	– deviation from Jan. mean	= Residual
5.298	– 4.090	– 1.138	= 0.070

The next step is to examine the autocorrelation in the residuals by plotting their correlogram. This correlogram is shown in Fig. 2A. Although there still appears to be some seasonality in the data, it is now much weaker and, more importantly, only one significant autocorrelation remains. This correlation occurs between adjacent months (i.e., at lag1) and can be removed by subtracting from the residual of each month, the residual of the previous month. The resulting differences are called lag1-transformations or first-order differences. Since it is not possible to compute a lag1 difference for the first month (i.e., January 1997), the lag1 transformation always results in the loss of the first data point. Figure 2B shows the correlogram of the lag1-transformed residuals and confirms that the transformation has removed the remaining autocorrelation from the data.

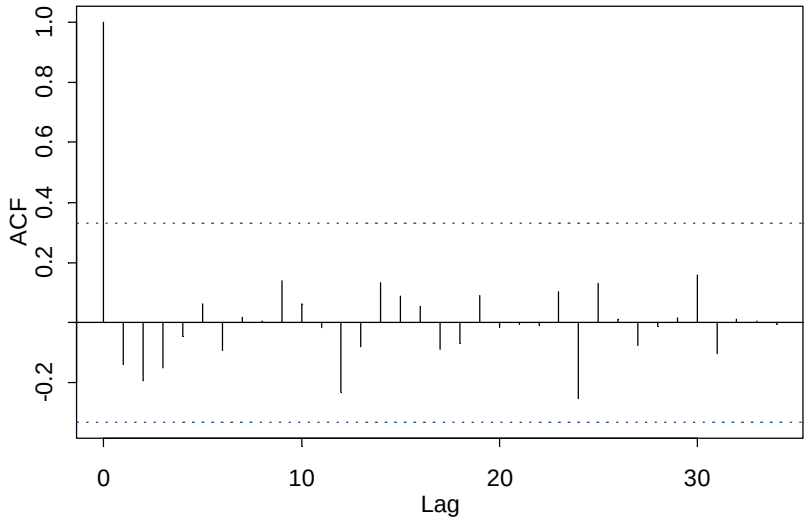
At this point, one has two choices. The first is to use an **ARIMA model** to test hypotheses about the time series. These models are extremely flexible and can be used to model seasonal time-series with autocorrelations at lags of 1 and beyond (Brockwell and Davis 1987; Diggle 1990). However, ARIMA models are mathematically complex and require the assistance of a skilled time series analyst. Fortunately, a much simpler approach is available for cases such as the present one in which the only significant autocorrelation occurs at lag1. This approach leads to the construction of an upper one-sided 95% confidence interval on the 3-year mean of the log-turbidity. The one-sided alternative hypothesis that the three-year median is less than 150 NTU can be evaluated by checking whether the upper bound on the back-transformed confidence

Fig. 2. Correlograms of seasonally-adjusted, monthly log-transformed turbidity data from the Mermantau River, January 1997- December 1999. (A) Data prior to lag1 differencing. (B) Data after lag1-differencing.

A. Correlogram of Seasonally-Adjusted 1997-1999 Data



B. Correlogram of Lag-1 Differenced 1997-1999 Data



interval is greater than 150. The log-scale upper bound is computed from the following formula:

$$UCL = \bar{y} + \left(t_{1-a,df} \times \frac{\sqrt{s_{adj}^2}}{\sqrt{nm}} \times \sqrt{\frac{1+\hat{f}_1}{1-\hat{f}_1}} \right) \quad (2)$$

where, $\bar{y}$ = the mean log-turbidity

$t_{1-a,df}$ = the t-statistic value associated with p=1-α and df degrees of freedom

n = the number of years of data

m = the number of months in a year

df = ((n×m)-m)/3

s_{adj}^2 = the variance of the seasonally adjusted log-turbidity

$\hat{f}_1$ = the estimated autocorrelation in the seasonally adjusted log-turbidity at lag1

The eleven steps required to obtain this confidence interval will now be illustrated with the data from Table 1.

1. Compute the square of each seasonally adjusted residual (i.e., square each value in column 5). The resulting squared values are given in Column 6.
2. Sum the squared residuals. The result is 8.348.
3. Compute the variance of the seasonally adjusted residuals:

   $$s_{adj}^2 = \frac{1}{(nm)-m} \times sum\ of\ squared\ residuals$$

   $= (1/((3\times12)-12)) \times 8.348$

   $= 0.348$

   and,

   $s_{adj} = \sqrt{0.348} = 0.590$
4. Multiply each residual by the value of the residual in the preceding month. These values are listed in column 7.
5. Sum the products from Step 4. The result is 3.955
6. Compute the autocorrelation between each log-turbidity and the log-turbidity in the preceding month by dividing the sum of the products of the lagged residuals by the sum of the squared seasonally adjusted residuals:

   $\hat{f}_1$ = 3.955/8.348 = 0.473

   note: this is the value of the lag1 autocorrelation shown in the correlogram in Fig. 2B
7. Compute the mean of the 36 monthly log-turbidity values (148.4/36). This value is 4.12.
8. Compute the degrees of freedom for the t-statistic = ((n×m)-m)/3 = ((3×12)-12)/3 = 8
9. Look up the value of the t-statistic associated with p=0.95 and df=8. The result is t=1.86.
10. Carry out the arithmetic in Eq. 2, substituting the values computed in steps 1-9,

    $$UCL = 4.12 + \left[1.86 \times \left(\frac{0.590}{6} \right) \times \sqrt{\frac{1+0.473}{1-0.473}} \right]$$

    $= 4.12 + (1.86 \times 0.098 \times 1.672)$

    $= 4.427$

11. Now back-transform the log-scale mean and the log-scale UCL to get the geometric mean and its upper 1-sided 95% confidence limit:
    Geometric Mean = Exp(4.12) = 61.68
    95% upper –sided confidence limit = exp(4.427) = 83.68

Having initially verified (Appendix C, Section C.2.5) that the 1980-2000 "population" of turbidity values fit a lognormal distribution, we can interpret the geometric mean and its 95% upper bound to be a valid estimate of the population upper 1-sided confidence interval on the population median. Since this interval does not include the criterion value, we can conclude that the sample evidence does not support the hypothesis that the mean/median turbidity in the Mermentau River was greater than or equal to 150 NTU during 1997-1999. Had we not log-transformed the data, our estimate of the mean would not have provided an unbiased estimate of the population median. Likewise, if we had not adjusted for autocorrelation, our estimate of the standard error of the geometric mean and its confidence interval also would have been biased. The procedure just described corrects for both problems without loss of data.

D.3 Nonparametric one-sample tests on means

Tests like the t-tests that require the sample data to have some specific parametric distributional form (e.g., the normal) are called **parametric tests.** When the data do not come from such a distribution, or when the data set is too small to tell what distribution it came from, or when a suitable normalizing transformation cannot be found, a class of tests called **nonparametric tests.** may be employed to compare the mean or median of a continuous water quality variable (e.g., turbidity) against a criterion value (e.g., the maximum allowable turbidity value in a stream). In this section we describe two such tests, the Wilcoxon signed ranks tests and Chen's modified t-test. These and other nonparametric tests useful for analyzing WQS data are described in Hollander and Wolfe 1999 and Millard and Neerchal 2001.

The formula for the exact Wilcoxon test statistic is described in Box 7. Like the t-test on log(X), the Wilcoxon signed-ranks test evaluates the null hypothesis that the median of the differences between the standard (M_0) and the sample values is zero. The sample value of W can be compared against the exact distribution of the Wilcoxon sign-ranks test statistic to determine the probability of obtaining a larger value under the null hypothesis (Hollander and Wolfe 1999). In practice, one usually uses statistical software (e.g., SAS; SPLUS EnvironmentalStats) to obtain the exact p-values.

Wilcoxon signed rank test depends on the following assumptions:

1. The variable must be measured on an interval scale; i.e. it can be either a count or a continuous random variable.
2. The sampling units from which the values of the variable were measured must be spatially and temporally independent
3. The frequency distribution of the variable values must be symmetric

Assumption 1 is verified simply by examining the measurement scale, assumption 2 is verified using the graphical methods illustrated in Sections C.2.4 and C.2.5, and assumption 3 is verified by constructing a frequency histogram of the sample data (e.g., Fig. 5a of Appendix C). When all the assumptions hold, the Wilcoxon signed ranks test is valid.

Box 7: Exact Wilcoxon Signed-Ranks Test for the Median

Assumptions

3. The population distribution of the attribute X is symmetric around the population median M.
4. The sample $X_1, X_2, \ldots, X_n$ is a random sample of *n* independent data points from the target population.

Let M_0 be the fixed criterion against which the population median M is compared. Consider each of the hypothesis pairs:

Case 1: $H_0 : M = M_0$ vs. $H_a : M > M_0$
Case 2: $H_0 : M = M_0$ vs. $H_a : M < M_0$

These are the steps for the exact Wilcoxon signed-ranks test.

Step 1. Select a significance level *a.*
Step 2. Calculate the difference $d_i = X_i - M_0$ for each of the *n* data points.
Step 3. Rank $[d_i]$, the absolute value of the differences, from lowest to highest. That is, the smallest $[d_i]$ will have rank 1, the second smallest is rank 2, …, the largest $[d_i]$ is rank *n*. If there are ties, assign the average of the ranks which would otherwise have been assigned to the tied observations. For instance, if 3 observations are tied for Rank 2 then assign the average of ranks 2, 3 and 4. Thus, these 3 tied observations each will have rank (2+3+4)÷3=3. The next rank after these 3 observations will be rank 5.

Step 4. Determine $sign(d_i)$, where $sign(d_i) = \begin{cases} +1 & \text{if } d_i \geq 0 \ (i.e., x_i \geq M_0) \\ -1 & \text{if } d_i < 0 \ (i.e., x_i < M_0) \end{cases}$.

Step 5. Calculate the sum W of the ranks with a positive sign.

$$W = \sum_{i=1}^{n} sign(d_i) \times Rank(|\, d_i \,|)$$

Step 6. Use the table of critical values of the Signed Ranks Statistic to find the critical value W_a. Compare *W* with W_a.
Case 1: If $W<W_a$ then reject H_0. Otherwise, accept H_0.
Case 2: If $W > \frac{n(n+1)}{2} - W_a$ then reject H_0. Otherwise, accept H_0.

Although the symmetry assumption of the Wilcoxon signed ranks is not as restrictive as the normality assumption, it precludes the analysis of skewed (e.g., lognormal) distributions, a serious drawback for analysis of environmental data. Chen's modified t-test offers a useful alternative for the analysis of skewed data. When the data are either right-skewed or left-skewed, the traditional t-test does not have good power (i.e., the Type II error rate is inflated).

Chen (1995) developed a modified form of the t-test that requires an estimate of the skew from which an adjustment is made to the value of the t-statistic to account for the skew. Chen's test requires the same assumptions as the traditional t-test (see Appendix C, Section C.3.1) with the notable exception that normality is *not* assumed. Computational details of Chen's modified t-test are summarized in Box 8a. Box 8b provides a detailed example the application of the test to evaluate the one-sided alternative hypothesis that the mean of a right-skewed distribution of herbicide concentrations is greater than the criterion value.

D.4 One-sample tests on binomial proportions

The ambient water quality criteria for several "conventional" parameters (e.g., dissolved oxygen, pH, temperature) are written in terms of the percentage of exceedant sampling units in a sample. For example, if the acceptable range of pH values for a body of water is set at 6.5-9.0, then any aliquot of water with a pH outside of this range will fail the standard. Consider a sample in which 17 of 100 aliquots collected from a lake have ph values outside this range; the sample exceedance rate is 17%. If the criterion for pH specifies that no more than 10% of the sampling units fail the standard, this sample would appear to exceed the criterion. However, a statistically valid assessment of the sample estimate requires that some accounting of the uncertainty in the sample be incorporated into the estimate. The statistical procedures described in this section and the next provide a method for doing so.

When the object of a study is to determine if the proportion of sampling units exceeding some water quality standard is greater than the proportion of exceedances permitted by a regulatory criterion (e.g., 10%) and the product of the sample size and the sample proportion ($n \times p$) and the product $n \times (1-p)$ are both ≥ 5.0, the proportions z-test may be used (Box 9a). In order to use a proportions test to evaluate a water quality criterion, the measured continuous values (e.g., pollutant concentrations) of the individual sample units must be converted to dichotomous (i.e., acceptable vs. exceedant) scores. Typically, this is done by creating a new variable, say Y, which is scored as 1 for each exceedant concentration or 0 for each acceptable concentration. The proportion of exceedances in the sample is then computed as the mean of Y. The one-sample binomial proportions test may be applied to evaluate two-sided alternatives (H_a population proportion $\neq$ criterion proportion) or lower (H_a population proportion $<$ criterion proportion) or upper (H_a population proportion $>$ criterion proportion) one-sided alternative hypotheses.

When the number of sampling units in the sample is large [i.e., $n \times p > 5.0$ and $n \times (1-p) \geq 5.0$] and the sampling units have been independently sampled from the target population, the test statistic in Box 9a will be normally distributed with mean $n \times p$ and variance $n \times p \times q$. For $\alpha=0.05$, a value of $z \geq 1.645$ will result in rejection of the null hypothesis that the population proportion

Box 8-a: Chen's Modified One-Sample t-Test for the Mean (One-sided Case)

Assumptions

5. X is a continuous random variable with population mean μ and variance s^2.
6. $X_1, X_2, \ldots, X_n$ is an independent sample of n individuals from the target population.

Let μ_0 be the fixed criterion against which the population mean μ is compared. Consider each of the 1-sided hypothesis pairs:

Case 1: $H_0 : \mu = \mu_0$ vs. $H_a : \mu > \mu_0$
Case 2: $H_0 : \mu = \mu_0$ vs. $H_a : \mu < \mu_0$

For either case, when the two assumptions hold, Chen's modified t-test for the null hypothesis vs. the alternative hypothesis can be performed as follows:

Step 1. Select the significance level *a*. (Typical values of *a* are 0.05 and 0.01.)
Step 2. Calculate the sample size, n, mean $\bar{x}$ and the sample variance s^2 (see Appendix C, Box 1-a).
Step 3. Compute the sample skewness:

$$\sqrt{\hat{b}_1} = \frac{\hat{m}_3}{\hat{s}^3}$$

Where: $\hat{m}_3 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (x_i - \bar{x})^3$

$$\hat{s}^3 = \left[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{3/2}$$

Step 4. Calculate the standard t-statistic.

$$t_0 = \frac{\bar{x} - m_0}{\sqrt{\frac{s^2}{n}}}$$

where $\bar{x}$ = the sample mean of the measured attribute, X
s^2 = the sample variance of the measured attribute, X
n = the number of sampling units in the sample
μ_0 = the fixed criterion against which the population mean is compared.

Box 8-a: Chen's Modified One-Sample t-Test for the Mean (Continued)

Step 5. Compute Chen's skew-adjusted t-statistic:

$$t_c = t_0 + a\left(1 + 2t_0^2\right) + 4a^2\left(t_0 + 2t_0^3\right)$$

Where: $a = \frac{\sqrt{\hat{b}_1}}{6\sqrt{n}}$

Step 6. Use the table of percentage points of student's t-distribution to find the value $t_{1-a,\ n-1}$ such that (1-*a*)x100% of the Student t distribution with n-1 degrees of freedom is less than $t_{1-a,\ n-1}$.

Step 7. Compare t_c with $t_{1-a,\ n-1}$.

Case 1: If $t_c > t_{1-a,\ n-1}$ then reject H_0.
If $t_c = t_{1-a,\ n-1}$ then reject H_0.

Case 2: If $t_c < t_{1-a,\ n-1}$ then reject H_0.
If $t_c = t_{1-a,\ n-1}$ then reject H_0.

Box 8-b: Example for Performing Chen's One-Sample t-Test for the Mean

Consider a random sample of 15 fish taken from a river. One measurement of Chlorophenoxy herbicide concentration in liver tissue was obtained from each fish. The sample values of the tissue concentrations and some intermediate calculations required for the sample mean, variance, and skewness estimates are:

HERBICIDE CONC.	DEVIATION	SUM OF SQUARED DEVIATIONS	SUM OF CUBED DEVIATIONS
10	-165.5	27,390	-4,533,086
13	-162.5	53,797	-8,824,102
20	-155.5	77,977	-12,584,131
36	-139.5	97,437	-15,298,836
41	-134.5	115,527	-17,731,974
59	-116.5	129,100	-19,313,142
67	-108.5	140,872	-20,590,431
110	-65.5	145,162	-20,871,442
110	-65.5	149,452	-21,152,453
136	-39.5	151,013	-21,214,083
140	-35.5	152,273	-21,258,822
160	-15.5	152,513	-21,262,546
200	24.5	153,113	-21,247,840
230	54.5	156,084	-21,085,961
1300	1124.5	1,420,584	1,400,844,570
2632			

The sample mean (175.5) and variance (101,470.29) are computed by dividing the sums of the first column and the last entry in the third columns by the sample size, n (15) and n-1, respectively. The values in column 2 are computed by subtracting the mean fro each value in column 1. Using the equations in Box 8-a, the skew can be computed in three steps:

Step 1. $$\hat{m}_3 = \left[\frac{15}{(15-1)(15-2)}\, x\, 1{,}400{,}844{,}570\right]$$
$$= 115{,}454{,}22.8$$

Step 2. $$\hat{s}^3 = (101{,}470.27)^{3/2}$$
$$= 32{,}322{,}752$$

Step 3. $$\sqrt{\hat{b}_1} = \frac{115{,}454{,}223}{32{,}322{,}743} = 3.572$$

Box 8-b: Example for Performing Chen's One-Sample t-Test for the Mean (Continued)

Step 4. Compute the value of a (see Box 8-a):

$$a = \frac{3.572}{6\sqrt{15}} = 0.15373$$

It is desired to test the null hypothesis that the mean tissue concentration of the herbicide is no more than 100 µg/kg vs. the alternative the mean is greater than 100 µg/kg/

$$H_0 : m \le 100 \quad \text{vs.} \quad H_a : m > 100$$

Step 5. Thus the value of the usual t-statistic is computer as:

$$t_0 = \frac{175.5 - 100}{\sqrt{101{,}470.27 \big/ 15}} = 0.918$$

Step 6. Compute Chen's modified t-statistic:

$$t_c = 0.918 + 0.15373(1 + 0.918^2) + \left[4 \times 0.15373^2 (0.918 + 2x0.918^3)\right]$$
$$= 1.563$$

Note: the difference between the values of t_c and t_0 reflect the adjustment for the skew in the sample data.

Step 7. Set the desired a-level, e.g., 0.05; thus for the 1-sided case and n=15 we need to find the critical value of associated with df=14 and 1-a=0.95 in the table of percentage points of student's t-distribution. This value is 1.761.

Step 8. Comparing $t_c = 1.563$ to $t_{0(14,095)} = 1.761$, we conclude that since the value of t_c is < t_0, the herbicide concentration in the sample of fish tissue support the null hypothesis that fish taken from the river attain the regulatory standard for chlorophenoxy herbicide.

Box 9-a: One-Sample Test for Proportions (Large Samples)

Assumptions

7. X is a dichotomous random variable taking on values 0, 1 denoting respectively the absence or presence of some attribute or condition (e.g., exceedance of some standard) observable for each sampling unit in the target population.
8. The sample $X_1, X_2, \ldots, X_n$ is a random sample of n sampling units from a population with a proportion, p , of its individuals with X=1
9. The values of n and p are such that both n×p and n× (1-p) = 5.0.

Let *p* denote the proportion of sampling units that exceed some criterion value. Let p_0 be the criterion value against which the population proportion p is compared. Consider each of the hypothesis pairs:

Case 1: $H_0 : p = p_0$ vs. $H_a : p > p_0$
Case 2: $H_0 : p = p_0$ vs. $H_a : p < p_0$

These are the steps for a large-sample (n>50) test for the population proportion.

Step 1. Select the significance level *a*. (Typically, *a* is 0.05 or 0.01.)
Step 2. Calculate the sample proportion *p* (see Box 1-a).
Step 3. Use the table of cumulative probabilities of the standard normal distribution to find the value $z_{1-a/2}$ such that (1-a)x100% of the standard normal distribution is below z_{1-a}.
Step 4. Calculate the test statistic z_c.

$$z_c = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where z_c = the standard normal test statistic
p = the proportion of exceedant sampling units in the sample
p_0 = the regulatory limit on the proportion of exceedances in the population.

Step 5. Compare z_c with z_{1-a}, the value of Z in the table of cumulative probabilities of the standard normal distribution that is associated with a probability = 1-*a*.
Case 1: If $z_c > z_{1-a}$ then reject H_0. Otherwise, accept H_0.
Case 2: If $z_c < z_{1-a}$ then reject H_0. Otherwise, accept H_0.

Box 9-b: Example for Performing a One-Sample Test for Proportions (Large Samples)

Consider a random sample of 100 monthly turbidity measurements from the Louisiana River between June 1980 and April 2000. The sample proportion of measurements that exceed 150 NTU is *p=0.19*. It is desired to test whether this sample proportion is indicative that the population proportion is larger than 0.15.

$$H_0 : p = 0.15 \quad vs. \quad H_a : p > 0.15$$

Step 1. The desired significance level is a=0.05.
Step 2. The sample proportion has been calculated to be *p=0.19*.
Step 3. The table of cumulative probabilities of the standard normal distribution yields $z_{0.95}$=1.645.
Step 4. Calculate the test statistic.

$$z = \frac{0.19 - 0.15}{\sqrt{\frac{(0.15)(0.85)}{100}}} = \frac{0.04}{\sqrt{0.001275}} = 1.12$$

Step 5. Since 1.2 < 1.645, then the null hypothesis cannot be rejected. Accept the null hypothesis and conclude that the true proportion of monthly turbidity measurements that exceed 150 NTU is less than or equal to 0.15.

of exceedances π is $\leq p_0$. Tables of the cumulative distribution of z are available in most elementary statistics texts and in all the standard commercial statistical software packages.

When the sample size criteria for the large sample normal approximation (Z-test) are not met, the behavior of the binomial becomes more discrete and is not well approximated by a continuous distribution such as the standard normal. For small samples, the exact binomial distribution must be used as the basis for both hypothesis tests and confidence intervals. The exact binomial test-statistic is simply the observed number of exceedances (r) out of the n sampling units in the sample. The p-value associated with r exceedances out of n sampling units is the upper-tailed cumulative binomial probability of observing r or more exceedances and can be computed with the formula in Box 10. The validity of the p-value for the exact binomial test rests on two assumptions: (1) the sampling units were selected from the target population by an independent, random process and (2) the underlying population exceedance probability is constant for every sampling unit in the population. Tables of the cumulative binomial P are available in many statistics texts (e.g., Hollander and Wolfe 1999) and from commercial statistical software packages (e.g., SAS; SPLUS EnvironmentalStats; StatExact).

The one-sided test of the null hypothesis that the observed proportion of exceedant sampling units comes from a population whose true proportion of exceedances is $\leq p_0$ is carried out by first finding the cumulative probability of observing $\geq$ r exceedances out of n sampling units (the p-value obtained from the equation in Box 10a). This p-value is then compared to the value of α specified by the analyst; if $p < \alpha$, she rejects H_0, otherwise she accepts H_0. Because of the discreteness of the binomial distribution, the actual α will generally be somewhat lower than the α specified in the DQOs. For example, if an investigator specifies $p_0 = 0.30$ and has a sample of n=17 with 11 exceedant sampling units and wants to test against the one-sided upper-tail alternative with $\alpha=0.05$, he will actually need to use $\alpha=0.0403$. This is because given n=17, the closest that one can actually come to p=0.05 (without exceeding it) is the cumulative probability of observing $\geq$ nine exceedances in the sample; i.e., p= 0.0403. This of course tells us that any number of exceedances ≥ 9 will cause rejection of H_0 for a sample size of n=17. The probability of observing ≥ 11 exceedances in a sample of 17 sampling units from a population with $p_0 = 0.30$ is 0.0032, which is $< \alpha=0.0403$; thus H_0 is rejected in this case.

D.5 Controlling Type I and II error rates for exact binomial tests

Although all of the interrelationships among n, power, and α described earlier hold for exact binomial tests, the relationships are complicated by the discreteness of the distribution. Table 2 summarizes these relationships for samples containing from 1 to 10 exceedances when the criterion level (p_0) is 10% exceedances, the sample size is fixed (n=10) and five different α-levels are specified. The following are the exact binomial probabilities of observing at least r exceedances out of a sample of 10 sampling units from a target population in which the true proportion of exceedances is 0.10:

Box 10-a: Exact Binomial Test for Proportions of Exceedances (Listing Case)

Binomial Distribution

Suppose there are *n* independent observations of a trait that has only two possible values (say *success* or *failure*). Let the probability *p* of observing a *success* be the same for all observations. Then the total number of successes, X, among the n observations has a binomial probability distribution, where the probability of observing *k* successes in a sample of size *n* is given by:

$$\Pr(X = k) = \left(\frac{n!}{k!(n-k)!} \right) p^k (1-p)^{n-k} \qquad \text{where } k = 0, 1, 2, 3, ..., n$$

To obtain the probability of observing a range of values for X, add up the probabilities of observing each of the values in the range. Tables of the cumulative probabilities are available in most statistics texts (e.g., Hollander and Wolfe, 1999) and from commercial statistics software packages (e.g., SAS, SPLUS EnvironmentalStats, StatExact).

Assumptions:

1. The sample is selected from the target population by an independent, random process.
2. The attribute of interest is exceedance of a specified criterion.
3. The underlying population probability of exceedance, p, is constant for every unit in the population.

Then the total number of exceedances in a random sample of size *n* is a binomial random variable.

In order to list a body of water as being impaired it ;has to be demonstrated that the population proportion of exceedances is greater than the regulatory limit. This leads to the null hypothesis that the true proportion of exceedances in the population is less than or equal to a standard maximum allowable proportion of exceedances, p_0. Following are the steps for performing the exact binomial test for the listing case.

Step 1. Specify the desired significance level a (usually 0.05 or 0.01).

Step 2. The exact binomial test statistic is the observed number of exceedances, *r*, in the sample.

Step 3. Calculate the p-value, *P*, the upper-tailed cumulative probability of observing at least *r* exceedances among *n* sampling units, when the true proportion is assumed to be p_0.

$$\Pr(X \geq r) = P = \sum_{k=r}^{n} \left(\frac{n!}{k!(n-k)!} \right) p_0^k (1-p_0)^{n-k}$$

where P = the upper-tailed cumulative binomial probability
p_0 = the regulatory limit on the proportion of exceedances in the population
n = the sample size
r = the observed number of exceedant sampling units in the sample.

Table of the cumulative probabilities are available in most statistics tests (e.g., Hollander and Wolfe, 1999) and from commercial statistics software packages (e.g., SAS, SPLUS, EnvironmentalStats, StatExact).

Step 4. Compare the p-value, *P*, with the desired significance level a.
If *P*<a then reject the null hypothesis; otherwise, accept the null hypothesis.

Box 10-b: Exact Binomial Test for Proportions of Exceedances (Listing Case)

Consider a random sample of 10 monthly turbidity measurements from the Mermentau River between June 1980 and April 2000. The measurements (in NTU) are: 34, 58, 87, 145, 14, 38, 62, 95, 160, 320. Suppose that the maximum allowable proportion of exceedances is 0.15. Based on the sample estimate (i.e., 2/10 -= 0.20) the proportion of exceedances, is the Mermentau River impaired with respect to the criterion that the proportion should be =0.15?

$$H_0 : p = 0.15 \quad vs. \quad H_a : p > 0.15$$

Step 1. Let the desired a=0.05.
Step 2. There are *r=2* exceedances in the sample.
Step 3. Calculate the p-value.

$$\begin{aligned}
\Pr(X \geq 2) = P &= \sum_{k=2}^{10} \left(\frac{10!}{k!(10-k)!} \right) (0.15)^k (0.85)^{10-k} \\
&= \left(\frac{10!}{2!(10-2)!} \right) (0.15)^2 (0.85)^{10-2} + \left(\frac{10!}{3!(10-3)!} \right) (0.15)^3 (0.85)^{10-3} \\
&\quad + \ldots + \left(\frac{10!}{10!(10-10)!} \right) (0.15)^{10} (0.85)^{10-10} \\
&= 0.2759 + 0.1298 + \ldots + (0.15)^{10} \\
&= 0.4557 .
\end{aligned}$$

Step 4. Since 0.4557 > 0.05, then the null hypothesis cannot be rejected. Accept the null hypothesis and conclude that the sample does not provide sufficient evidence that the Mermentau River is impaired.

Table 2. Exact Binomial probabilities, and Type II error rates for n=10 and 5 different prespecified Type I error rates for listing water bodies.

SAMPLE SIZE	NUMBER OF EXCEEDANCES	SPECIFIED TYPE I	ACTUAL TYPE 1	TYPE II ERROR	LOWER 1-SIDED EXACT 95% CI	MIN. NO. TO REJECT
10	1	0.05	0.0128	0.9872	0.100 (0.005, 1.000)	4
		0.20	0.0702	0.9298	0.100 (0.022, 1.000)	3
		0.25	0.0702	0.9298	0.100 (0.028, 1.000)	3
		0.30	0.2639	0.7361	0.100 (0.035, 1.000)	2
		0.35	0.2639	0.7361	0.100 (0.042, 1.000)	2
10	2	0.05	0.0128	0.8791	0.200 (0.037, 1.000)	4
		0.20	0.0702	0.6778	0.200 (0.083, 1.000)	3
		0.25	0.0702	0.6778	0.200 (0.096, 1.000)	3
		0.30	0.2639	0.3758	0.200 (0.109, 1.000)*	2
		0.35	0.2639	0.3758	0.200 (0.122, 1.000)*	2
10	3	0.05	0.0128	0.6496	0.300 (0.087, 1.000)	4
		0.20	0.0702	0.3828	0.300 (0.158, 1.000)*	3
		0.25	0.0702	0.3828	0.300 (0.176, 1.000)*	3
		0.30	0.2639	0.1493	0.300 (0.193, 1.000)*	2
		0.35	0.2639	0.1493	0.300 (0.209, 1.000)*	2
10	4	0.05	0.0128	0.3823	0.400 (0.150, 1.000)*	4
		0.20	0.0702	0.1673	0.400 (0.239, 1.000)*	3
		0.25	0.0702	0.1673	0.400 (0.261, 1.000)*	3
		0.30	0.2639	0.0464	0.400 (0.281, 1.000)*	2
		0.35	0.2639	0.0464	0.400 (0.300, 1.000)*	2
10	5	0.05	0.0128	0.1719	0.500 (0.222, 1.000)*	4
		0.20	0.0702	0.0547	0.500 (0.327, 1.000)*	3
		0.25	0.0702	0.0547	0.500 (0.351, 1.000)*	3
		0.30	0.2639	0.0107	0.500 (0.373, 1.000)*	2
		0.35	0.2639	0.0107	0.500 (0.393, 1.000)*	2
10	6	0.05	0.0128	0.0548	0.600 (0.304, 1.000)*	4
		0.20	0.0702	0.0123	0.600 (0.419, 1.000)*	3
		0.25	0.0702	0.0123	0.600 (0.445, 1.000)*	3
		0.30	0.2639	0.0017	0.600 (0.468, 1.000)*	2
		0.35	0.2639	0.0017	0.600 (0.489, 1.000)*	2
10	7	0.05	0.0128	0.0106	0.700 (0.393, 1.000)*	4
		0.20	0.0702	0.0016	0.700 (0.516, 1.000)*	3
		0.25	0.0702	0.0016	0.700 (0.542, 1.000)*	3
		0.30	0.2639	0.0001	0.700 (0.566, 1.000)*	2
		0.35	0.2639	0.0001	0.700 (0.587, 1.000)*	2

* PROPORTION OF EXCEEDANCES SIGNIFICANTLY > 0.10

Table 2. (continued)

SAMPLE SIZE	NUMBER OF EXCEEDANCES	SPECIFIED TYPE I	ACTUAL TYPE 1	TYPE II ERROR	LOWER 1-SIDED EXACT 95% CI	MIN. NO. TO REJECT
10	8	0.05	0.0128	0.0009	0.800 (0.493, 1.000)*	4
		0.20	0.0702	0.0001	0.800 (0.619, 1.000)*	3
		0.25	0.0702	0.0001	0.800 (0.645, 1.000)*	3
		0.30	0.2639	0.0000	0.800 (0.667, 1.000)*	2
		0.35	0.2639	0.0000	0.800 (0.687, 1.000)*	2
10	9	0.05	0.0128	0.0000	0.900 (0.606, 1.000)*	4
		0.20	0.0702	0.0000	0.900 (0.729, 1.000)*	3
		0.25	0.0702	0.0000	0.900 (0.753, 1.000)*	3
		0.30	0.2639	0.0000	0.900 (0.773, 1.000)*	2
		0.35	0.2639	0.0000	0.900 (0.791, 1.000)*	2
10	10	0.05	0.0128	0.0000	1.000 (0.741, 1.000)*	4
		0.20	0.0702	0.0000	1.000 (0.851, 1.000)*	3
		0.25	0.0702	0.0000	1.000 (0.871, 1.000)*	3
		0.30	0.2639	0.0000	1.000 (0.887, 1.000)*	2
		0.35	0.2639	0.0000	1.000 (0.900, 1.000)*	2

* PROPORTION OF EXCEEDANCES SIGNIFICANTLY > 0.10

r	Pr(no. exceedances ≥ r)
1	0.6513
2	0.2639
3	0.0702
4	0.0128
5	0.0016

Thus if an investigator wants to specify α=0.05, he has to consider an actual α that is slightly larger (0.0702) or slightly smaller (0.0128). This is because there are only 10 possible exceedance values >0 for a sample of size 10. The conservative choice will always be to choose the probability closest to, but less than, the desired α. The desired α-levels are specified in the third column of Table 2 and the corresponding actual α-levels are in column four. These values are fixed for all samples of n=10, regardless of how many exceedances (column two) they might contain. Note that since at least 4 exceedances are required to achieve a probability ≤ an actual α =0.0128, only samples with ≥ four exceedances will result in rejection of H_0. Similarly, at least three exceedances will be required for rejection at specified α-levels of 0.20 and 0.25, while a minimum of two exceedances will be required to reject at α-levels of 0.30 and 0.35. Because the numbers of exceedances that are required to reject the null hypothesis in a sample of ten sampling units (column seven) are specific to the α-level, they do not change with the number of exceedances actually observed in the sample.

The observed proportion of exceedances in a sample (r/n) has a strong effect on the Type II error probability. The difference between the observed proportion and the criterion value (p_0=0.10) is actually an effect size measure. Recall that for fixed sample size, α-level, and variance, the effect size will determine the observed power and hence the observed Type II error probability. The variance of a binomial response is [p ×(1-p)], thus the variance changes with the number of exceedances in the sample. However, for proportions in the range of 0.20 to 0.80, the differences in the variance are quite small. Therefore most of the change in Type II error for a given α-level in Table 2 is due to the increasing effect size. For example, the Type II error rate for the specified α=0.05 in a sample with 3 exceedances is 0.6496, but in a sample with 4 exceedances it is only 0.3823; the corresponding sample variances are 0.21 and 0.24. Note in Table 2 that the Type II error rates (for all α-levels) drop precipitously between r=1 and r=5 exceedances. This corresponds to the gray region in Figure 16 and indicates that a sample of size of 10 has reasonably good power when the effect size is ≥ 0.40 (i.e., $\delta \geq 5/10$ – the criterion value). In other words, our telescope (sample with n=10 and α=0.0128) does not have sufficient power for us to distinguish clearly (i.e., with Type II error < 0.20) two mountains (test statistic distributions) whose centers (proportions of exceedances) are any closer than 0.40.

The lower one-sided 100(1-α)% exact binomial confidence intervals (column 6) provide an estimate of the proportion of the exceedances in the target water body. Whenever the lower bound on this estimate is > 0.10, we reject H_0 and conclude that the sample evidence supports the alternative hypothesis that the water exceeds the acceptable standard. All such confidence intervals are marked with an asterisk. Notice that although the criterion is 0.10, when the specified α-level is 0.05, samples with 10%, 20% and 30% exceedances do not have lower bounds > 10%. This reflects the uncertainty in the estimate, but it also clearly demonstrates the high likelihood of erroneously accepting that such bodies of water meet the water quality

standard when, in fact, they do not. The minimum number of exceedances required for a particular 100(1-α)% lower bound to be > than 10% is exactly the same as the minimum number required by the exact binomial test to reject H_0 for a specified α-level. For example, the first 95% confidence interval with an asterisk occurs when there are four exceedances in the sample. Similarly, the first 70% confidence interval with an asterisk occurs when there are only two exceedances. Thus regardless of which statistical tool one chooses (i.e., confidence intervals or exact binomial tests), the decision remains the same. Furthermore, both tools are subject to the same Type I and II error rate problems when small sample sizes are used.

The exact binomial test may also be employed for the complementary process of delisting a body of water that was previously found to be impaired (Box 11a). Using the same criterion level (i.e., the population exceedance rate must be < 10%), a table similar to Table 2 may be constructed to illustrate the delisting problem. Table 3 summarizes relationships among n, α, β for the delisting scenario for sample sizes of 22-28. Regardless of the specified α, it will never be possible to delist a previously listed water with a sample size less than 22. Recall that in the listing case,

$$H_0 : p \leq p_0 \quad vs. \quad H_a : p > p_0$$

Where p = observed proportion of exceedances

p_0 = standard maximum allowable proportion of exceedances

Thus, we reject the null hypothesis and list any water whose proportion of exceedances > 0.10 (p_0). A statistician would say that the rejection region for this decision scenario is between 0.10 and 1.0. By contrast, for the delisting case, the null and the alternative hypotheses for the listing scenario are essentially flip-flopped:

$$H_0 : p \geq p_0 \quad vs. \quad H_a : p < p_0$$

Therefore, we will only delist a listed a body of water if the proportion of exceedances in the sample is < 0.10. The statistical rejection region in this case is much smaller (0-0.10) than for the listing case (0.10-1.0). This requires a much more powerful telescope; i.e. much larger sample sizes. Consequently, it will always be much more difficult to delist a body of water than it was to list it in the first place.

The third column from the left in Table 3 is the compliment of the Type I error rate (i.e., 1-α); column three is the Type II error rate (β), and the last column is the maximum number of exceedances out of a sample of n sampling units which can be tolerated if the water is to be delisted. The fourth column is just the value in the last column divided by n and is therefore the maximum tolerable exceedance rate for a sample of given n and α. Comparing Tables 2 and 3, generally speaking, there is a much greater likelihood that a "bad", previously unlisted water, will not be listed than that a "bad" previously listed water will be delisted.
In the context of water quality attainment decisions, both Type I and Type II errors are cause for concern. For the landowner, an erroneous listing of the water (Type I error) on his property may result in severe financial hardship. On the other hand, the failure to list and subsequently restrict the use of a impaired water (Type II error) may have dire public health consequences. Therefore, it is difficult to argue that either type of error is more important than the other. This would seem to justify the specification of balanced α- and β-levels. Unfortunately, as we have demonstrated

Box 11-a: Exact Binomial Test for Proportions of Exceedances (Delisting Case)

Assumptions

10. The sample is selected from the target population by an independent, random process.
11. The attribute of interest is exceedance of a specified criterion.
12. The underlying population probability of exceedance, p, is constant for every unit in the population.

Then the total number of exceedances in a random sample size of *n* is a binomial random variable.

The exact binomial test may also be employed for the complementary process of delisting a body of water that has been previously found to e noncompliant. In this case, the null hypothesis that the true proportion of exceedances in the population is greater than or equal to a standard maximum allowable proportion of exceedances, p_0. The steps for performing the test are similar to the listing case.

Step 1. Determine the significance level *a* (usually 0.05 or 0.01.)

Step 2. The exact binomial test statistic is the observed number of exceedances, *r*, in the sample.

Step 3. Calculate the p-value, *P*, the *lower-tailed* cumulative probability of observing *r* or fewer exceedances among *n* sampling units, when the true proportion is assumed to be p_0.

$$\Pr(X \le r) = P = \sum_{k=0}^{r} \left(\frac{n!}{k!(n-k)!} \right) p_0^k (1-p_0)^{n-k}$$

where P = the lower-tailed cumulative binomial probability
p_0 = the standard maximum allowable proportion of exceedances in the population
n = the sample size
r = the observed number of exceedant sampling units in the sample.

Step 4. Compare the p-value, *P*, with the desired significance level *a*.
If $P<a$ then reject the null hypothesis. Otherwise, accept the alternative hypothesis.

TABLE 3. ERROR RATES FOR SAMPLES CONTAINING THE MAXIMUM NUMBER OF ALLOWABLE EXCEEDANCES TO DELIST FOR SAMPLE SIZES 22-28

N	SPECIFIED TYPE I ERROR RATE	PROB. KEEPING BAD WATER ON LIST (1-α)*	PROB. KEEPING GOOD WATER ON ON LIST (β)	OBSERVED EXCEEDANCE	MAX. NO. EXCEEDANCES TO DELIST
22	0.05	0.9015	0.6406	0.0000	0
	0.20	0.9015	0.6406	0.0000	0
	0.25	0.9015	0.6406	0.0000	0
	0.30	0.9015	0.6406	0.0000	0
	0.35	0.6608	0.2641	0.0455	1
23	0.05	0.9114	0.6406	0.0000	0
	0.20	0.9114	0.6406	0.0000	0
	0.25	0.9114	0.6406	0.0000	0
	0.30	0.9114	0.6406	0.0000	0
	0.35	0.6849	0.2642	0.0435	1
24	0.05	0.9202	0.6399	0.0000	0
	0.20	0.9202	0.6399	0.0000	0
	0.25	0.9202	0.6399	0.0000	0
	0.30	0.7075	0.2642	0.0417	1
	0.35	0.7075	0.2642	0.0417	1
25	0.05	0.9282	0.6396	0.0000	0
	0.20	0.9282	0.6396	0.0000	0
	0.25	0.9282	0.6396	0.0000	0
	0.30	0.7288	0.2642	0.0400	1
	0.35	0.7288	0.2642	0.0400	1
26	0.05	0.9354	0.6393	0.0000	0
	0.20	0.9354	0.6393	0.0000	0
	0.25	0.9354	0.6393	0.0000	0
	0.30	0.7487	0.2642	0.0385	1
	0.35	0.7487	0.2642	0.0385	1
27	0.05	0.9419	0.6391	0.0000	0
	0.20	0.9419	0.6391	0.0000	0
	0.25	0.7674	0.2642	0.0370	1
	0.30	0.7674	0.2642	0.0370	1
	0.35	0.7674	0.2642	0.0370	1
28	0.05	0.9477	0.6388	0.0000	0
	0.20	0.9477	0.6388	0.0000	0
	0.25	0.7849	0.2642	0.0357	1
	0.30	0.7849	0.2642	0.0357	1
	0.35	0.7849	0.2642	0.0357	1

* FOR N <29 IT IS NOT POSSIBLE FOR ACTUAL α TO BE $\leq$ 0.05

previously, simultaneous control of α and β to levels ≈ 0.05 requires unrealistically large sample sizes.

It has become standard practice in the scientific literature to specify an α-level of 0.05; in those papers where it is considered, the maximum acceptable β-level is generally set at 0.20. Freedman et al. (1991) provide the following explanation of the origin of the 0.05 α-level:

> "R. A. Fisher was the first to use such tables [i.e., tables of test statistics associated with probabilities of 0.05 and 0.01] and it seems to have been his idea to lay them out this way. There is a limited amount of room on a page. Once the number of α-levels was limited to [two values], 5% and 1% stood out as nice round numbers and they soon acquired a magical life of their own. With computers everywhere, this type of table is obsolete. So are the 5% and 1% α-levels."

Smith et al. (2001) have suggested balancing the error rates at moderate levels (e.g., ≤ 0.15). This requires the investigator to specify both α and β levels, *a priori*, at step 6 of the DQO process. Because of the discrete nature of the binomial distribution, α and β levels can only be specified subject to the specification of a minimum number of exceedances required to reject H_0 (what Smith et al. call the "cut-off"). In the present example H_0: $p = 0.10$. As explained previously, the Type II error rate cannot be set without first specifying a minimum effect size. Smith et al. propose that a population exceedance rate of 0.25, "indicates severe problems and represents the minimum violation [rate] we would almost always want to detect". Thus they recommend specifying $p=0.25$ for the population under H_a, which is equivalent to specifying a minimum effect size of 0.15. Employing these specifications, a table of balanced Type I and Type II error rates for sample sizes less than 50 can be computed (Table 4).

Due to the discreteness of the binomial, the error rates can only be balanced for one sample size per each unique cut-off value (Table 4, column 2). The error rates for sample sizes less than 28 are probably too high to be acceptable to most regulators. As with so many investigations of sample size, power and precision, the results in Table 4 seem to suggest than n=30 is the "magic number". Clearly, the use of balanced error rates does not solve the problems associated with small sample sizes; however, it does remove the inequality associated with protection of one error rate at the expense of the other. Moreover, the balanced error presentation clearly reveals the risks associated with making decisions based on small sample sizes.

A similar table of balanced error rates can be computed for the delisting scenario (Table 5). Although the criterion for listing is 10% exceedance, we will adopt the reasoning of Smith et al. (2001) and specify that the exceedance rate should be $\geq 20\%$ (i.e., H_0: $P \geq 20\%$) to justify keeping a body of water on the list; whereas, any body of water with <10% exceedance (i.e., H_a: $P < 10\%$) will be taken off the list. As pointed out earlier, the mathematics involved with flipping the null and the alternative hypotheses result in a very small rejection region, so that it will be hard to reject the null hypothesis except when we have large samples. The practical consequence of all this is that much more evidence is required to delist than to list. The results in Table 5 demonstrate that if the minimally acceptable balanced error rates are to be held to $\leq 15\%$, the minimum sample size for delisting a body of water must be 59 (i.e., slightly more than double

Table 4. Balanced Type I and II error rates for the exact binomial for eight sample sizes for listing water bodies.

SAMPLE SIZE	MIN. NO. TO REJECT (CUTOFF)	TYPE I ERROR PROB	TYPE II ERROR PROB	POWER(%)
4	1	0.34	0.32	68.4
10	2	0.26	0.24	75.6
16	3	0.21	0.20	80.3
22	4	0.17	0.16	83.8
28	5	0.14	0.14	86.5
34	6	0.12	0.11	88.6
40	7	0.10	0.10	90.4
46	8	0.08	0.08	91.8

Table 5. Balanced Type I and II error rates for the exact binomial for thirteen sample sizes for delisting water bodies.

SAMPLE SIZE	MAX. NO. TO REJECT (CUTOFF)	TYPE I ERROR PROB	TYPE II ERROR PROB	POWER(%)
25	3	0.23	0.24	0.7636
32	4	0.20	0.21	0.7885
39	5	0.18	0.19	0.8097
46	6	0.16	0.17	0.8281
52	7	0.16	0.14	0.8560
59	8	0.14	0.13	0.8690
66	9	0.12	0.12	0.8800
73	10	0.11	0.11	0.8900
80	11	0.10	0.10	0.9000
87	12	0.09	0.09	0.9080
94	13	0.08	0.08	0.9150
100	14	0.08	0.07	0.9270
108	15	0.07	0.07	0.9290

the number to needed list with the same error rates). An example of the computation of exact binomial proportions for the delisting case is presented in Box 11b.

With respect to employing exact binomial procedures to support water quality attainment decisions, the preceding discussion indicates the following:

1. The sample of sampling units should be obtained from a design that insures independent and representative sampling of a target population clearly bounded in space and time.
2. Tests with balanced Type I and II errors are preferable to tests designed primarily to minimize Type I error rates.
3. Balanced Type I and Type II error rates should be less than 0.15.
4. The minimum effect size employed to compute balanced error rates should be based on careful consideration of the sampling costs and consequences of the Type II error.
5. In order to meet the requirements of 1 and 3, the sample should contain at least 28 sampling units for the listing case and 59 for the delisting case.

6. The binomial test and associated confidence intervals will be valid only if the exceedance rate is constant throughout the target population; i.e., there should be any spatial or temporal heterogeneity in the distribution of the exceedances.

The binomial model will lead to appropriate decisions only to the extent that the sample of n sampling units represents the true spatial and temporal variability of the body of water in question. The binomial model is no panacea for inadequate sample size, poor (i.e. nonrandom or restricted) sampling designs, or populations with extreme heterogeneity of exceedance rates. For heterogeneous populations, it will probably be more prudent to base water quality attainment decisions on tests comparing sample means or medians against the appropriate pollutant concentration or biological abundance criterion.

D.6. Estimation of the total exceedances in a 3-year period

Both acute and the chronic water quality criteria may be exceeded once per 3-year assessment period without consequence; however, waters with two or more exceedances of either standard will be listed as non-attainment waters. This is equivalent to requiring the total number of exceedances to be less than 2 for any given 3-year assessment period. Because 3-year attainment criteria are based on sampling from local bodies of water within the 3-year period, estimates of the total number of exceedances are subject to sampling error and other sources of uncertainty. Thus, it is necessary to quantify this error so that the appropriate confidence intervals and hypothesis tests can be constructed. A strategy for doing so is presented in this section.

We will take as an example 36 monthly sampling units collected from a particular river reach during 1997 –1999 for the assessment of the acute selenite criterion. The selenite concentration in each monthly sampling unit is compared against the selenite CMC (186 µg/l) and scored 1 if it exceeds the CMC; zero, otherwise. The sum of these scores is the sample estimate of the total number of exceedances in the population. But what is the population? Recall that the CMC is based on the hourly mean of repeated measurements taken within a 24-hour period. The total

Box 11-b: Exact Binomial Test for Proportions of Exceedances (Delisting Case)

Consider the example in Box 10-b. Suppose instead that the Mermentau River has been listed as noncompliant. Suppose also that the maximum allowable proportion of exceedances is 0.25. Based on the sample proportion of exceedance of 0.2, can the Mermentau River be deemed compliant and delisted?

$$H_0 : p = 0.25 \quad vs. \quad H_0 : p < 0.25$$

Step 1. Let a=0.05.
Step 2. There are *r=2* exceedances in the sample.
Step 3. Calculate the p-value.

$$\begin{aligned}
\Pr(X \le 2) = P &= \sum_{k=0}^{2} \left(\frac{10!}{k!(10-k)!} \right) (0.25)^k (0.75)^{10-k} \\
&= \left(\frac{10!}{0!(10-0)!} \right) (0.25)^0 (0.75)^{10-0} + \left(\frac{10!}{1!(10-1)!} \right) (0.25)(0.75)^{10-1} \\
&\quad + \left(\frac{10!}{10!(10-2)!} \right) (0.25)^2 (0.75)^{10-2} \\
&= 0.0563 + 0.1877 + 0.2816 \\
&= 0.5256 .
\end{aligned}$$

Step 4. Since 0.5256 > 0.05 then the null hypothesis cannot be rejected. Accept the null hypothesis and conclude that the Mermentau River cannot be delisted.

number of such means that could be computed for a 3-year period is just the number of days in 3 years ≈ 3×365 = 1095 days. Thus, the total number of exceedances in the population must be between zero and 1095. Since we are actually categorizing the day upon which a particular monthly assessment was made as an exceedant or a compliant day, it is the individual days of the 3-year period that are the sampling units for the estimates of the total numbers of exceedant days.

Based on these definitions of the population (i.e., 1095 days) and the sample (36 days randomly sampled out of the target population), and the dichotomous outcome (the daily mean is classified as exceeding/not exceeding the CMC) it might at first appear that this is simply a problem in binomial proportions in which the sample proportion of exceedant days is compared to the criterion value of 2/1095 (i.e., proportion of CMC exceedances must be < 0.0018). However, there is one crucial difference; whereas, the target populations in the previous applications of the binomial distribution were essentially infinite (e.g., the number of 1-liter aliquots of water flowing through a river reach in a month), in this case, we have a finite target population. The hypergeometric distribution is the appropriate probability model for estimation of dichotomous proportions from a finite population (Cochran 1977; Thompson 1992). The probability of observing x number of exceedances in a sample of size n, given that it was randomly selected from a target population of fixed size N, containing exactly k exceedances is,

$$\Pr(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} = \frac{\dfrac{k!}{x!(k-x)!}\,\dfrac{(N-k)!}{(n-x)!(N-k-n+x)!}}{\dfrac{N!}{n!(N-n)!}} \tag{3}$$

Where, x = the number of exceedant days in the sample of n days
n = the number of days sampled
N = the total number days in a 3-year assessment period = 1095
k = the number of exceedant days in the population

A related expression provides an estimate of the probability of observing ≤ x exceedant days in the sample. The expression is just the sum of all of the individual probabilities of observing 0 to x exceedances, each calculated with equation 3. This expression is called the cumulative distribution function (CDF),

$$\Pr(X \le x) = \sum_{r=0}^{x} \frac{\binom{k}{r}\binom{N-k}{n-r}}{\binom{N}{n}} \tag{4}$$

The probability of observing more than x exceedances (i.e., = x+1) is just 1 minus CDF. This upper tailed hypergeometric probability provides the basis for deciding between the following two hypotheses:

H_0: the number of exceedances in the target population = k

H_a: the number of exceedances in the target population > k.

Or equivalently:

H_0: the proportion of exceedances in the target population = k/N

H_a: the proportion of exceedances in the target population > k/N.

For any given x, the resulting upper tailed p-value is compared to α and if it is smaller, the null hypothesis is rejected; otherwise it is accepted. Here, we want to choose x such that our specified a rate is not exceeded. This is the finite sampling counterpart to the exact binomial test for binary outcomes from an infinite population. Like the binomial, exact $100\times(1-\alpha)\%$ confidence intervals can be formed for the total number of exceedances in the population or the corresponding proportion of exceedances. The lower 1-sided exact confidence limit on the number of exceedant days in the population is computed as the smallest k such that,

$$\Pr(X \ge x) = 1 - \sum_{r=0}^{x-1} \frac{\binom{k}{r}\binom{N-k}{n-r}}{\binom{N}{n}} > a \tag{5}$$

and the upper 1-sided exact confidence limit is computed as the largest k such that,

$$\Pr(X \le x) = \sum_{r=0}^{x} \frac{\binom{k}{r}\binom{N-k}{n-r}}{\binom{N}{n}} > a \tag{6}$$

The values of k that satisfy equations 5 and 6 are the lower and upper bounds on the total number of exceedant days during the 3-year monitoring period. To obtain the corresponding bounds on the proportion of exceedant days in the 3-year period, we simply divide each of the k values by N (i.e., 1095). Two-sided confidence intervals can be formed by replacing α with $\alpha/2$ in equations 5 and 6 (Buonaccorsi 1987; Wendell and Schmee 2001).

As an example, consider the selenite monitoring program in which samples were collected monthly for 3 years. The CMC criteria require that there be no more than 1 CMC exceedance by any of the 1095 daily means during that period. What is the probability of obtaining 0, 1, or = 2 exceedant days in the sample of 36 days, if the true population had only 1 exceedant day out of 1095? We can compute the first two probabilities by substituting x=0 and x=1 into equation 3, specifying n=36, N=1095 and k=1; the third probability (which must be 0) is just one minus the sum of the first two (i.e., Equation 4):

$$\Pr(X=0)=\frac{\dfrac{1!}{0!(1-0)!}\ \dfrac{(1095-1)!}{(36-0)!(1095-1-36+0)!}}{\dfrac{1095!}{36!(1095-36)!}}=0.968$$

$$\Pr(X=1)=\frac{\dfrac{1!}{1!(1-1)!}\ \dfrac{(1095-1)!}{(36-1)!(1095-1-36+1)!}}{\dfrac{1095!}{36!(1095-36)!}}=0.032$$

$$\Pr(X\le 1)=0.968+0.032=1.000 \Rightarrow \Pr(X\ge 2)=1-(0.968+0.032)=0.0$$

Therefore, the 1-sided upper-tailed p-value is zero. This is intuitive; if the target population truly contains only 1 exceedant day, then the probability that there will be more than one exceedance in any sample drawn from it will be zero. Thus if we find even 1 exceedance in a sample of 36, we must reject the null hypothesis that there was only 1 exceedance in the target population during the 3-year monitoring period. This result is more understandable if we think in terms of the proportion of exceedances. A sample with 1/36 exceedant days implies a population with an exceedance rate of 0.0278. Since there are 1095 days in the population, the expected number of exceedant days is 0.0278×1095˜ =31 days.

Using Equations 5 and 6, we can obtain confidence intervals for the total number exceedant of days. In this case we would like to compare the lower 1-sided 95% confidence limit to the hypothesized value of 1 exceedant day, so we use Eq. 5 and find that the lower bound is 2 exceedant days. This is greater than the hypothesized value (1) so it agrees with the exact p-value; however, we should note that the lower one-sided confidence interval on the estimate is extremely small relative to the estimate itself (i.e., 31), suggesting that the sample size of n=36 yields very low power. We can further illustrate this problem by comparing the probabilities of obtaining 0 exceedance from a population with only 1 exceedance (p=0.968) to the probability of obtaining 0 exceedances for a population with 20 exceedances (i.e., k=20). We obtain the later probability as,

$$\Pr(X=0)=\frac{\dfrac{20!}{0!(20-0)!}\ \dfrac{(1095-20)!}{(36-1)!(1095-20-36+1)!}}{\dfrac{1095!}{36!(1095-36)!}}=0.509$$

Thus, even if the true population has as many as 20 exceedant days there is still a 50-50 chance of obtaining 0 exceedant days in a sample of 36 days. This means that finding 0/36 exceedances does not provide a great deal of assurance that the true population does not contain substantially more than 1 exceedant day; i.e., the false negative error rate with n=36 is quite high (0.935) and the statistical power is low (0.065). A sample with zero out of 36 exceedances will not provide reliable evidence for attainment of the CMC standard.

Because the once-in-3-years criterion specifies that 2 or more exceedances in a year is a violation of the standard, if one assumes that there is no measurement error in the determination of the concentration of the pollutant, then any sample with 2 or more exceedant days will automatically indicate (i.e., without the need for statistical testing or estimation) that the target population had at least 2 exceedant days during the 3-year monitoring period. Thus the only situations where hypothesis testing and/or confidence interval estimation will be necessary is when there are only 0 or 1 exceedant days in the sample. Due to relationships described above, the occurrence of 1 exceedant day in a sample of n = 245 days will indicate rejection of the null hypothesis at the a=0.05 level.

Hence for any feasible sample size (e.g., 36 monthly assessments), the occurrence of even one exceedance will lead to a decision to list the body of water. This may appear to be excessively conservative at first; however, given sample sizes on the order of 36, most bodies of water with 1 or even 2 exceedances in a 3-year period *will not be listed*. This is because for a body of water that actually experiences 2 days during which a pollutant concentration exceeds the criterion, the probability that a sample of 36 assessments will contain at least one of the two exceedances is only 6.4%. Consequently a sample size of 36 will be associated with a false negative rate (i.e., Type II error rate) of 93.6%. Moreover, the probability that a sample of 36 assessments will contain both exceedances is only 0.1%.

The problem of determining the minimum sample size required for exact hypergeometric tests that maintain simultaneously specified false positive and false negative error rates is tedious without the aide of statistical software. At present, only one readily available statistical software package (PASS) will compute power and sample size for the exact hypergeometric distribution. Thus it is more difficult to apply the DQO procedure to the once-in-3-years CMC criterion based on the exact hypergeometric estimates, than for some of the other distributions discussed in this appendix (e.g., chi-square or t).

For either the acute or the chronic criteria, the Type I (false positive) and Type II (false negative) error rates and the power of the exact hypergeometric test against the upper 1-sided alternative for samples that contain 1 exceedant day can be specified as follows:

First, find the largest value of x_0 (i.e., the largest number of exceedant days) such that:

$$\Pr\left[x \geq x_0 \,|k, n, N \right] \leq a \tag{7}$$

That is, we specify a critical value x_0 for the number of exceedant days, so that given the hypothesized number of exceedant days in the 3-year period (k), our sample size, and the population size, the probability of observing a larger number of exceedant days, is less than or equal to the alpha-level. For the situation under consideration that number is 1 day.

Next, having specified a value for x_0, we calculate the power and Type II error rate as:

$$\begin{aligned} &Power\,|k = \Pr\left[x \geq x_0 \,|n, k, N \right] = 1 - \Pr\left[x < x_0 \,|n, k, N \right] \\ &\Rightarrow False\ Negative\ Error\ Probability\,|k = \Pr\left[x < x_0 \,|n, k, N \right] \end{aligned} \tag{8}$$

Thus, for specified N and a population that is assumed to have the smallest number of exceedances that will lead to a non-attainment decision (i.e., k=2), we can find the minimum sample size as the smallest value of n, in Eq. 8, that will yield a Power = 1-ß (where ß = the

largest acceptable false positive error rate specified in the DQO). For the selenite example, to achieve a power of at least 0.85, the minimum sample size is n=1010. By contrast, when a sample of only 36 days is used, the power is only 0.064.

The same approach can be used to make inferences about the total number of chronic assessments in a 3-year period. Of course, chronic assessments are based on 4-day means, so there are only 1095/4 = 274 possible chronic assessments that could be made in any one 3-year period. Computations for confidence intervals and exact p-values can be made using equations 3 –6, substituting 274 for all values of N. Employing Eq. 4 to obtain the CDF for the case of x=1 exceedant mean, we find that the p-value for the corresponding test against the upper one-sided alternative is 0.13, indicating support for H_0. The estimated total number of chronic exceedances in the 3-year population, based on a sample with 1 chronic exceedance out 36 monthly assessments, is 8 exceedant days with a lower 95% bound of 1 exceedant day. Since the lower bound is within the acceptance region (i.e., it is = 1 exceedant day), a sample of only one exceedance out of 36 sampling units offers support for the null hypothesis that the body of water attains the chronic CMC standard; however, the probability of a false negative decision error with n=36 and a population size of N=274 is 0.9832. Thus a sample of n=36 that has only 0 or 1 exceedant days from a population that actually has 2 exceedant days, is almost certain lead to an incorrect attainment decision; sample sizes this small will not meet the DQO requirement that false negative error rates be less than 0.15. In fact the minimum sample size required for a false negative error rate = 0.15 is n=253.

The preceding calculations and discussion assume that exceedances occur independently of one another during the 3-year assessment period. However, it is more likely that exceedances will be associated with transient episodic events of variable duration. For example, consider a single discharge of flyash that elevates selenite concentrations above the CMC for 20 consecutive days in a reach of river during a 3-year assessment period. If the reach were being monitored on a regular monthly basis, there would be a 20/30.4 (66%) probability that a single exceedance would be recorded, a 34% probability that no exceedances would be recorded, and a 0.0% probability that two or more exceedances would be recorded during the 20-day episode.

The difficulties associated with the once-in-3-years assessments occur because the regulation allows 1 extremely rare event (e.g., 1 exceedant day out of 1095 or 1 out of 274 days), but not 2 extremely rare events. Thus, the width of the gray region is only 1/1095 for acute criteria or 1/274 for chronic criteria. When criteria are based on the assessment of such rare events, the false negative decision error rates become inflated to very high levels unless samples sizes are increased so that n ˜ N.

D.7 References

Altman, M.J., D. Machin, T. Bryant, and M. Gardner. 2000. Statistics with Confidence (2nd Ed.). BMJ Books, London, UK.

Beyer, W.H. (Ed.). 1971. CRC Basic Statistical Tables. CRC, Cleveland, Ohio.

Buonaccorsi, J. 1987. A note on confidence intervals for proportions in finite populations. American Statistician 41(3):215-218.

Brockwell, P.J. and R. A. Davis. 1987. Time Series: Theory and Methods. Springer-Verlag. New York.

Cleveland, W.S. 1993. Visualizing Data. Hobart Press. Summit, New Jersey, USA.

Cochran, W.G. 1977. Sampling Techniques (3rd ed.). New York, Wiley.

Diggle, Peter. 1990. Time series: a biostatistical introduction. Oxford University Press. Oxford, UK.

Freedman, D., R. Pisani and R. Purves. 1997. Statistics 3rd Ed. W.W. Norton & Co. New York

Hollander, M. and D. A. Wolfe. 1999. Nonparametric Statistical Methods, 2nd Ed. J. Wiley & Sons. New York.

Johnson, N. L. and S. Kotz.. 1969. Discrete Distributions. Houghton Mifflin. New York City, NY.

Lindley, D. and W. F. Scott. 1995. New Cambridge Statistical Tables (2nd Ed.) Cambridge University Press, New York. NY.

Millard, S. P. and N. K. Neerchal. 2000. Environmental Statistics with S-PLUS. CRC Press, Boca Raton, FL.

O'Brien, R.G. 1998. A tour of UnifyPow: a SAS module/macro for sample size analysis. Proceedings of the 23rd SUGI Conference, 1346-1355. SAS Institute, Cary, NC.

Smith, E. P., K. Ye, C. Hughes, and L. Shabman. 2001. Statistical assessment of violations of water quality standards under Section 303(d) of the Clean Water Act. Environ. Sci. Technol. 35:606-612.

Steel, G.D., J.H. Torrie and D.A. Dickey. 1996. Principles and procedures of statistics: a biometrical approach. McGraw-Hill, New York.

Stephan, C.E., D.I. Mount, D.J. Hansen, J.H. Gentile, G.A. Chapman, and W.A. Brungs. 1985. Guidelines for Deriving Numerical National Water Quality Criteria for the Protection of Aquatic Organisms and their Uses: EPA.

Thompson, S.K. 1992. Sampling. J. Wiley and Sons. New York, New York.

Wendell, J. and J. Schmee. 2001. Likelihood confidence intervals for proportions in finite populations. American Statistician 55(1):55-61.

D.8 Glossary

alternative hypothesis - In a statistical hypothesis test there are two competing hypothesis, one of which, the alternative hypothesis, describes the set of conditions complementary to those described under the null hypothesis. For example: if the null hypothesis states that the mean pH of a set of samples is less than or equal to 5.0, the alternative hypothesis must be that the mean pH is greater than 5.0.

ANOVA Model - an acronym for Analysis of Variance. ANOVA models are linear models in which the total variance in a response is partitioned into two components: one due to treatments (and possible interactions among them) and the other due to random variability. In the simplest case where there is only one treatment factor, if the treatments have no effect on the response, the ratio of the variance components should be close to 1.0. If the treatments effect the response means, the ratio of the treatment component to the random component will be greater than one. Under the null hypothesis that the treatments have no effect, the sampling distribution of the ratio of the two variance components, each divided by their respective **degrees of freedom**, will be an F-distribution.

ARIMA Model - an acronym for autoregressive integrated moving average model. ARIMA models are linear models for regression and/or discrete treatment effects, measured through time, on responses that have been differenced at the appropriate **lag** distances.

autocorrelation - the internal correlation of a set of measurements taken over time and/or space. The correlation arises from the fact that points closer together in space and/or time tend to be more alike than those that are further apart. The autocorrelation function (either spatial or temporal) is a mathematical expression that relates the strength of the correlation to the distance (called the **lag**) between measurements.

Bayesian statistical inference - An approach to **inference** or **estimation** in which a process (e.g., a random binomial process) is proposed for the generation of a set of data. A mathematical model called a **likelihood** is specified for the process, such that the model parameters are random variables. A distribution, called the prior distribution, is developed for the parameters based on what is known about them, prior to collection of the data. Data are then collected and a mathematical principle called Bayes theorem is used to derive a second distribution of the parameters, called the posterior distribution, from the data and the prior distribution. The appropriate inference is then obtained from the posterior distribution. The Bayesian approach differs from the the classical frequentist approach in that is utilizes the investigator's prior knowledge of the system through the prior distribution.

bias - the systematic or persistent distortion of a measurement process that causes errors in one direction.

binary characteristic - a characteristic that can only have two possible values.

census - a study that involves the observation and/or measurement of every member of a population.

confidence interval - a range of values, calculated from the sample observations, that is believed, with a particular probability, to contain the true population parameter value. For example, a 95% confidence interval implies that if the estimation process were repeated again and again, then 95% of the calculated intervals would be expected to contain the true parameter value.

confidence level (also called the the confidence coefficient) - the probability that the confidence interval will include the true parameter value; Equivalently, 1-the probability (α)that the true value is *not* contained within the interval .

continuous random variable - A random variable which may take on an infinite number of values.

convenience sample - a sample collected from a target population without implementation of a probability-based design. Sampling units are selected based on ease of data collection, without clear reference to an underlying frame; e.g., the collection of water samples near bridges rather than randomly throughout the stream reach to which an inference is desired. Because many (perhaps the majority) of the population sampling units have little or no probability of selection to the sample and because sample coverage typically is restricted to some potentially homogeneous subset of the target population, data from convenience samples are not valid for statistical inference to the target population.

correlation coefficient – A scale-invariant measure of the association between 2 variables that takes on values between –1.0 and +1.0. The correlation coefficient has a value of plus one whenever an increase in the value of one variable is accompanied by an increase in the other, zero when there is no relationship (i.e., the 2 variables are independent of one another), and minus one (-1) when there is an exact inverse relationship between them.

correlogram - a plot or graph of the sample values of the autocorrelation coefficient of a time series against different values of the **lag**.

decision error - an error that occurs when data misleads an investigator into choosing the wrong response action, in the sense that a different action would have been taken if the investigator had access to unlimited “perfect data” or absolute truth. In a statistical test, decision errors are labeled as false rejection (Type I) or false acceptance (Type II) of a null hypothesis.

degrees of freedom (df) - As used in statistics, df has several interpretations. A sample of n values is said to have n degrees of freedom, but if k functions of the sample values are held constant, the number of degrees of freedom is reduced by k. In this case, the number of degrees of freedom is conceptually the number of independent observations in the sample, given that k functions are held constant. By extension, the distribution of a statistic based on n independent observations is said to have $n-p$ degrees of freedom, where p is the number of parameters of the distribution.

discrete random variable - A random variable which may take on only a finite number of values.

dispersion - the amount by which a set of observations are spread out from their mean and/or median.

effect size - In a **one-sample test**, the difference between the sample mean and a pre-specified criterion or standard value. In a **two-sample test**, the effect size is the expected difference between the mean of a treatment group or ambient site vs. the mean of a control group or reference site. Associated statistical tests typically evaluate the null hypothesis of a zero effect size vs. the alternative that the effect size is nonzero.

effective sample size - When data are collected from cluster-correlated populations, there is redundancy in the information carried by more highly correlated individuals. Thus, correlated individuals carry less information than do uncorrelated individuals. The effective sample size is the number of uncorrelated individuals from a simple random sample that would carry information equivalent to the information in the sample of correlated individuals. The effective sample size is always less than the apparent sample size; how much less, is a function of the strength of the correlation and the sampling design that was used to collect the data.

estimation - the process of providing a numerical value for a population parameter on the basis of information collected from a sample.

experimental design - the arrangement or set of instructions used to randomize subjects to specific treatment or control groups in an experimental study. Such a procedure generally insures that results are not confounded with other factors and thus provides scientifically defensible inferences regarding causal effects of the treatments.

exploratory data analysis (EDA) - an approach to data analysis that may reveal structure and/or relationships among measured or observed variables in a data set. EDA emphasizes informal graphical procedures that typically are not based on prior assumptions about the structure of the data or on formal models.

extreme values - the largest and smallest values (and perhaps their neighboring values) among a sample of observations.

frequentist statistical inference - an approach to statistics based on the likelihood of an observed result in a large or infinite number of independent repetitions of the same sampling or experimental procedure (e.g., see the frequentist definition of the **confidence interval** in this glossary).

geometric mean - a measure of central tendency calculated by back-transforming the mean of a set of log-transformed observations. If the original data come from a log-normal distribution, the sample geometric mean will provide an unbiased estimate of the sample **median**.

heterogeneous - a term denoting inequality or dissimilarity of some quantity of interest (most often a variance) in a number of different groups, populations, etc.

homogeneous - a term denoting equality or similarity of some quantity of interest (most often a variance) in a number of different groups, populations, etc.

imprecision/precision - A term describing the degree of spread among successive estimates of a population parameter by a sample statistic. The standard error of a sample estimator (e.g., the standard error of the mean) is a measure of imprecision/precision in the estimator. A high degree of spread (imprecision) will lead to an increased likelihood of a decision error, while a reduction in spread will lead to a corresponding reduction in the likelihood of a decision error. Generally, precision will be increased by increasing the sample size.

independence - essentially, two events are said to be independent if knowing the outcome of one tells us nothing about the outcome of the other. More formally, two events *A* and *B* are said to be independent if Probability (A and B) = Prob(A) × Prob(B).

inference - the process of drawing conclusions about a population on the basis of measurements or observations made on a sample of individuals from the population.

lag - the distance, in units of time or space, between two events or locations. For example, an event occurring at time $t+k$ $(k>0)$ is said to lag behind the event occurring at time t, by an amount of time equal to lag k.

likelihood - the probability of a set of observed data, given the value of some parameter or set of parameters associated with a model for the underlying process that is hypnotized to have generated the data. For example, if we obtain 9 heads in 10 tosses of a coin, the likelihood of observing this result, given that the coin is fair (i.e., the binomial parameter p=0.50), is approximately 0.0098.

log-transformation - a transformation on a variable, X, obtained as, Y=ln(X) or Y=ln(x+c), where c is a constant positive value (e.g., 1.0). This transformation is useful for normalizing continuous variables with skewed distributions and/or stabilizing the variance of a variable whose standard deviation increases as a function of its mean, prior to statistical analysis with tests (e.g., t-tests) that assume normality and/or variance homogeneity..

maximum likelihood - a procedure for estimating the value of a parameter(s) of a model of the underlying process that produced some particular set of observations, such that the resulting estimate maximizes the likelihood of the observed data. For example, the maximum likelihood estimate for the binomial parameter P, given an experiment in which one obtains 9 heads in 10 tosses is P=0.90. The likelihood of obtaining 9 heads given an underlying binomial process with P=0.90, is 0.3874. Note that the estimate P=0.90 leads to a much larger likelihood than an estimate of P=0.50 does (0.0098; see definition of **likelihood**). In fact there is no value of P that will yield a larger likelihood of obtaining 9 heads out of 10 tosses than the estimate P=0.90; thus, P=0.90 is the maximum likelihood estimator of P.

median - in a sample or a population, the median is the value of a random variable such that half of the sampling units have larger values and half have smaller values. When the population or sample size is 2N+1, the median is the value of the random variable associated with the $N+1^{th}$

ordered sampling unit; when the population or sample size is 2N, the median is average of random variable values of the sampling units with ranks N and N+1. If the population is normal the median will equal the mean. If the population is log-normal the median will equal the **geometric mean**.

Monte Carlo methods - methods for finding solutions to mathematical and statistical problems by simulation. Often used when the analytic solution of the problem is intractable, or when real data are difficult to obtain, or to evaluate the behavior of statistics or models under a variety of hypothetical conditions which may or may not be directly observable in nature.

nonparametric statistical methods (also called distribution-free methods) - Statistical techniques of **estimation** and **inference** are often based on the assumption of some underlying parametric process; for example, one that generates responses that are normally distributed. By contrast, nonparametric estimation and testing procedures do not depend on the existence of an underlying parametric process. Consequently, nonparametric techniques are valid under relatively general assumptions about the underlying population. Often such methods involve only the ranks of the observations rather than the observations themselves.

noncentral t-distribution - the expected distribution of the t statistic when the alternative hypothesis is true. This contrasts with central t-distribution (usually referred to simply as the "t-distribution") which is the expected distribution of the t-statistic when the null hypothesis is true. In general, the probability that an observed t-statistic comes from a non-central t-distribution will be large (e.g., $P>0.20$) when the probability of that it comes from a central t-distribution is low (e.g., $P<0.001$), and vice versa.

null hypothesis - a hypthesis about some presumed prevailing condition, usually associated with a statement of "no difference" or "no association" (see also **alternative hypothesis**).

one-sample tests - Statistical tests that evaluate the null hypothesis that there is no difference between a sample statistic (e.g., mean or proportion) and a fixed criterion or standard value.

parametric continuous distribution - the probability distribution of a continuous random variable, specified by a mathematical function of the population parameters; e.g., the normal distribution with parameters, μ and σ^2.

parametric statistical methods - tests and estimation procedures that depend on the complete specification of an underlying parametric probability distribution of the population from which the sample was drawn. The estimators and test statistics that are based on functions of the estimates of the population parameters under the assumed population distribution model (e.g. normal) are valid only if the assumed population model is valid. An example is the t-statistic which assumes an underlying normal population.

percentiles - the set of divisions of a set of data that produce exactly 100 equal parts in a series of values.

population - any finite or infinite collection of "units" that completely encompasses the set individuals of interest. In environmental studies, populations are usually bounded in space and time; e.g., the population of smallmouth bass in Leech Lake, Minnesota on July 1, 2000.

population parameter - a constant term(s) in a mathematic expression, such as a probability density function, that specifies the distribution of individual values in the population. Parameters typically control the location of the center of the distribution (location parameters), the spread of the distribution (scale or dispersion parameters) and various aspects of the shape (shape parameters) of the distribution (see also: **probability density function**).

power of a statistical test -the probability of rejecting the null hypothesis when it is false. Notice that we would like always to reject a false hypothesis; thus, statistical tests with high power (i.e., power >0.80) are desirable. Generally the power of a test increases with the number of individuals in the sample from which the test was computed.

precision - a term applied to the uncertainty in the estimate of a parameter. Measures of the precision of an estimate include its standard error and the confidence interval. Decreasing the value of either leads to increased precision of the estimator.

probability-based sample - a sample selected in such a manner that the probability of being included in the sample is known for every unit on the sampling frame. Strictly speaking, formal statistical inference is valid only for data that were collected in a probability sample.

probability density function (PDF)- for a continuous variable, a curve described by a mathematical formula which specifies, by way of areas under the curve, the probability that a variable falls within a particular range of values. For example, the normal probability density function of the continuous random variable X, is:

$$\frac{1}{s\sqrt{2p}}\exp\left[-\left(\frac{1}{2s^2}\right)(x-m)^2\right]$$

The normal probability density function has two **parameters**, the mean and variance , μ and σ^2. The mean is the location parameter and the variance is the scale parameter; the normal distribution does not have any shape parameters. The graph of the normal probability density function is the familiar "bell curve".

rank - the relative position of a sample value within a sample.

relative frequency - the frequency of occurrence of a given type of individual or member of a group, expressed as a proportion of the total number of individuals in the population or sample that contains the groups. For example, the relative frequencies of 14 bass, 6 bluegill, and 10 catfish in a sample of 30 fish are, respectively: 46.7%, 20.0% and 33.3%.

representative sample - A sample which captures the essence of the population from which it was drawn; one which is typical with respect to the characteristics of interest, regardless of the manner in which it was chosen. While representativeness in this sense cannot be completely

assured, probability-based samples are more likely to be representative than are judgement or convenience samples. This is true because only in probability sampling will every population element have a known probability of selection.

sample - a set of units or elements selected from a larger population, typically to be used for making inferences regarding that population.

sampling design - a protocol for the collection of samples from a population, wherein the number, type, location (spatial or temporal) and manner of selection of the units to be measured is specified.

sampling distribution - the expected probability distribution of the values of a statistic that have been calculated from a large number of random samples. For example, the sampling distribution of the ratios of each of the means from 100 samples (each with n=30) to their respective variances will be a t-distribution with 29 degrees of freedom.

sampling error - the difference between a sample estimate and the true population parameter due to random variability in the composition of the sample vs. that of the target population.

sampling frame - the list from which a sample of units or elements is selected.

sampling unit - the members of a population that may be selected for sampling.

significance level (α) - the level of probability at which it is agreed that the null hypothesis will be rejected; α is also the probability of a Type I error.

skewness - a measure of the asymmetry in a distribution, relative to its mean. A right-skewed distribution is composed mostly of small values lying close to the mean but possesses a few values that are much larger than the mean. Conversely, a left-skewed distribution is composed mostly of values lying close to the mean but possesses a few values that are much smaller than the mean.

square-root transformation - a transformation on a variable, X, obtained as, Y= or Y=$\sqrt{X + 1/2}$. This transformation is useful for normalizing a discrete variable with a Poisson distribution and/or stabilizing the variance of a variable whose variance is proportional to its mean, prior to statistical analysis with tests (e.g., t-tests) that assume normality and/or variance homogeneity.

standard deviation - the square root of the **variance**.

standard error - the standard error of a sample statistic, θ, (say a sample mean or proportion) is the standard deviation of the values of that statistic computed from repeated sampling of the target population, using the same sampling design (e.g., stratified simple random sampling) and the same sample size, *n*. For example, the standard error of the mean is the sample **standard deviation**/n.

standard normal distribution - a normal distribution whose mean is 0 and whose variance is 1.

statistic - a quantity calculated from the values in a sample (e.g., the sample mean or sample variance)**.**

statistical distribution - a probability distribution used to describe a statistic, a set of observations or a population.

statistical test of hypotheses - a statistical procedure for determining if a sample provides sufficient evidence to reject one statement regarding the population of interest (the null hypothesis) in favor of an alternative statement (the **alternative hypothesis**).

target population - the set of all units or elements, bounded in space and time, about which a sample is intended to produce inferences.

two-sample tests - Statistical tests that evaluate the null hypothesis that there is no difference between a sample statistic (e.g., mean or proportion) in a treatment group or at an ambient monitoring site and the sample statistic in a control group or at a reference site.

Type I error (α) - the error that occurs when a decision maker rejects the null hypothesis when it is actually true. Also called the false rejection decision error, or false positive decision error.

Type II error (β) - the error that occurs when a decision maker accepts a null hypothesis when it is actually false. This is also called the false acceptance decision error, or false negative decision error. The power of a statistical test is 1-β.

variance (population) - the variance of a finite population of N values - $x_1, x_2, \ldots x_N$ - is simply the average of the squared difference between the individual observations and the population mean.

variance (sample) - the variance of n sample observations is simply the average of the squared differences between the individual observations and the sample mean, divided by (n-1).

variogram - a plot of the sample values of the variance of a spatially referenced variable vs. the corresponding lag distances

Summary of June 2000 Stakeholder Meeting on the Consolidated Assessment and Listing Methodology

Overview

This document summarizes four stakeholder meetings hosted by the U.S. Environmental Protection Agency (EPA) on June 27 and 28, 2000, in Washington, D.C. at the Disabled American Veterans' Hall. EPA's Office of Water called the meetings in early June to obtain input from key water quality management stakeholders regarding the Assessment and Watershed Protection Division's (AWPD's) plans and initial ideas for developing a Consolidated Assessment and Listing Methodology (CALM) to identify impaired waters under Section 303(d) and prepare water quality assessment reports under Section 305(b) of the Clean Water Act.

General themes emerging from the meeting are identified first followed by separate summaries of each of the four meetings. The meetings were designed to obtain input from stakeholders as follows:

Meeting 1: Point Source Representatives
Meeting 2: Non-Governmental Organizations
Meeting 3: Nonpoint Source Representatives
Meeting 4: State Agency Representatives

At each meeting, each participant was given an opportunity to speak so that EPA could consider all points of view. This document therefore captures the diverse comments of individual participants and cannot be construed as representing a consensus of the participants or any group of them.

Participants' observations were recorded on flipcharts during the meetings. During and at the end of each meeting, participants were asked to review the flipcharts for accuracy and completeness. With minor changes to improve clarity for those who were not present, the detailed summaries of the separate meetings contain the original flipchart language. Participants' comments are not necessarily in the order they were offered at the meeting. They are organized in part to correspond to key questions identified during an introductory session at the outset of each meeting by Margarete Heber, Chief of the Monitoring Branch in AWPD. [A copy of the slides Ms. Heber used during her presentation is included at the end of this summary as Appendix A (PowerPoint format, 71kb).]

Agenda

Each meeting followed the same basic agenda:

1. Introduction of participants and review of meeting objectives;
2. Brief presentation on the Consolidated Assessment and Listing Methodology by Margarete Heber;

3. Facilitated discussion between participants and EPA staff, focusing on participants' views;
4. Wrap-up/review of next steps; and
5. Opportunity for comment by non-participant observers.

The meetings were facilitated by Ross & Associates Environmental Consulting, Ltd. Each participant was specifically invited to speak during the meeting and almost all did so. The meetings were facilitated to assure fair opportunity for all to speak, to help keep to the schedule, and to help stay on subject. However, most of the discussion was informal. From time to time, EPA staff answered questions or asked questions of the participants, but during most of the meeting, EPA representatives listened to the statements of the participants.

Participation

Participation was generally limited to representatives of the interests for which each meeting was organized (point source representatives at Meeting 1, non-governmental organization representatives at Meeting 2, etc.). Participants generally decided independently which of the four meetings to attend. In a few cases, participants in a meeting may have represented interests other than those targeted for that session. (For example, a representative of a public water supply system participated in Meeting 3.)

Participants were seated at a table, along with EPA representatives. Also participating in each meeting was Mr. Tom Morrisey, Co-Chair of the Monitoring Committee of the Association of State and Interstate Water Pollution Control Authorities (ASIWPCA). Each meeting was open to the public and several observers were present at all meetings. The number of participants from the targeted stakeholder group for each meeting varied considerably, due in part to the relatively short notice for the meeting.

Meeting 1: 25 point source representatives;
Meeting 2: 4 non-governmental organization representatives;
Meeting 3: 9 nonpoint source representatives; and
Meeting 4: 21 state agency representatives.

Over the course of the two days of meetings, eighty-two individuals attended, including EPA staff, observers, and meeting facilitators. Attendees are listed in Appendix B of this document.

CALM Follow-up Schedule

EPA's schedule for proceeding on the CALM initiative has changed since the June 27-28 meetings. Please check the EPA website for the latest information.

Themes Emerging from the Four Meetings

General Reaction to EPA's Plans to Develop Guidance on a Consolidated Approach

- Most participants **favored** the concept of developing a consolidated approach.
- Several participants noted that Sections 303(d) and 305(b) of the Act have **different purposes** that may require different approaches. For example, Section 303(d) listing has regulatory implications whereas Section 305(b) reports are a snapshot of water quality conditions. Section 303(d) requires more rigorous determinations. Some participants, however, express concern about differences in monitoring approaches for 303(d) and 305(b).
- Several participants noted that EPA needs to "raise the bar" on the **quality of data** used, especially in the Section 303(d) listing process.
- Several participants suggested that EPA's guidance **should not be unduly prescriptive** but should allow for a variety of good State methodologies. Some participants suggested that the guidance should set forth questions for States to answer about their assessment and listing processes and provide examples of acceptable and outstanding responses.
- Some participants expressed support for consolidating 303(d) lists and 305(b) reports on a **four-year cycle**.
- Some participants said that EPA should take steps to ensure greater **consistency** among States and between the CWA and Safe Drinking Water Act both in conducting assessments and making attainment determinations; others said that consistency within each State is paramount. Some State participants expressed concern about inconsistency among EPA regional offices.

<u>Determining Attainment/Nonattainment</u>

- Participants **supported** EPA's plan to provide **guidance** on how to define/determine attainment and impairment.
- Many participants said EPA's guidance should **allow for State or regional variation** in methods for determining attainment, but that EPA should be more specific than it is currently about what types of data are appropriate for making any attainment/impairment determination.
- While participants did not agree on how the issues should be addressed, the areas most often suggested as **needing guidance** were:
    - The **quantity and quality of data** needed to make a determination;
    - **Sampling frequency and the number of exceedances** that would constitute nonattainment;
    - How to incorporate **multiple data sets**;
    - How to handle **historical/aged data**;
    - Data requirements and process for **303(d) delisting**;
    - Proper uses of **evaluated data** generally;
    - How/whether to use **fish advisories** (especially under 303(d));
    - When/whether to use **inferences** from land use data or reference similar water bodies;
    - When/whether to use **biological data** to determine attainment status;
    - Use of **effluent data** from sources to predict attainment/nonattainment;
    - Interpretation, calibration and validation of **water quality models**;
    - Assessing and determining attainment status of **interstate water bodies**;

    - Dealing with **uncertainty**;
    - Defining **"pollution" and "pollutant"** for purposes of 303(d) listing decisions; and
    - The use of **"independent application"** of various water quality parameters to determine attainment status.
- Some participants said that EPA should issue guidance on determining attainment of designated uses, **narrative criteria**, wetlands criteria, aquatic life criteria, and/or antidegradation provisions; others said that water quality standards which are unclear or non-numeric need further clarification through rulemaking (e.g., establishment of "translator" provisions, revision, or replacement) before they can serve as a basis for attainment determinations. Some participants said that States need flexibility to interpret their non-numeric standards and that EPA should not over-emphasize numeric criteria.
- Most participants said that States should establish quality assurance/quality control parameters that could be followed by stakeholders willing to perform monitoring and that any data, including **stakeholder data**, should be used without bias by States in making attainment determinations, as long as the data are collected in accord with the prescribed parameters.
- Many participants noted that it is important for State and EPA attainment determinations to be as **transparent to the public** as possible and to be open about data limitations; several suggested that a consensus State/stakeholder process should be used to develop assessment and attainment determination protocols and to decide how 303(d) lists should be organized/prioritized. Some participants suggested that EPA should define "transparency" in its guidance and that the guidance should establish better ways of communicating 303(d) listing decisions and their significance for sources, stakeholders and the public.
- Several participants noted that States need more and better **expertise and expert assistance** in the proper interpretation of models and the conduct and uses of statistical analyses. Some suggested that EPA and State biologists should engage in more direct communication to enhance consistency and provide for technology transfer.
- Some participants said that **listing and delisting** decisions should be based on the same **degree of scientific rigor**, while others suggested that some questionable data indicating impairment should be considered determinative until better data is collected.

<u>Comprehensive Assessments</u>

- There was **general support** for the concept of assuring that all waters are monitored, but there was considerable concern about whether States have (or would ever have) the **capacity** to monitor all waters.
- Some participants suggested that, through the CALM process, EPA can **encourage more monitoring** and provide incentives to States to perform more monitoring.
- Many participants discussed the use of **probabilistic monitoring** approaches (the necessity to use such approaches given resource limitations, the need for guidance

on their use, and concern about whether they are appropriate for use in making attainment determinations especially for 303(d) purposes). Many State representatives supported use of probabilistic approaches for some purposes.

- Many participants expressed concern about the **definition/designation of water quality segments** and suggested that guidance is needed on the amount and quality of data required for geo-spatial determinations and what constitutes "representative" data for making water segment decisions.
- Several participants urged EPA not to jeopardize the **5-year rotating basin** approach used in a number of States.
- Several participants, including some point source representatives, expressed an interest in **point sources' conducting more instream monitoring** to supplement State efforts. Some others were concerned about requiring permittees to perform instream monitoring because the costs and burdens could be significant. Several participants agreed that if a point source has demonstrated stable compliance through effluent monitoring, effluent testing requirements could be reduced in exchange for instream monitoring.
- Some participants suggested that States should establish **Monitoring Councils** to assure proper coordination among the various agencies and entities that conduct monitoring. Some State participants expressed concern about their ability to coordinate effectively with the large number of interested/affected stakeholders.

For more specific information on views expressed by participants, please read on to the separate meeting note, below.

Meeting 1 Notes:
Point Source Representatives

General Comments

- Point source representatives are interested in a consolidated approach; appreciate the opportunity to provide input to CALM process.
- Supportive of CALM because it provides a mechanism to use locally collected data.
- There are differences in regulatory outcomes/consequences that flow from Section 305 and 303 actions; this needs to be recognized.
- EPA should raise the bar on quality of data used in Section 303(d) de-listing.
- Don't let current lack of data influence how to develop guidance (but, approach needs to be do-able by States). Need to balance certainty, resource and information needs.
- Approach should be less prescriptive for States; emphasize watershed approach (rotating basins).
- Guidance should cover how to conduct data review, with trigger points for Section 303(d) listing.
- States should give more emphasis and place greater importance on Section 305(b) reports to the public.

- We are concerned about timing. Will the CALM initiative lag behind the TMDL process?
- We are concerned about adequacy of resources (e.g., to run water quality models).
- Some States have good listing methodologies: EPA should not prescribe one way but should use State methodologies as templates.
- EPA's CALM schedule does not accommodate State rule-making process requirements.

Attainment Decisions

General Comments

- We need certainty, clear direction re: content, defensible data and how to incorporate multiple data sets.
- The process needs "integrity". Don't assume waters previously listed under Section 303(d) should continue to be listed under the new CALM methodology unless impairment is shown under the new methodology.
- States should develop an "I don't know" list of waters for which we don't have enough information to list under Section 303(d); make these waters a priority for additional monitoring.
- When States identify causes and sources of nonattainment, too much is assumed. This information has not been based on field data; "not all data is equal".

Is a water quality standard attained/not attained?

- Guidance needs to define impairment (is it present day, worst case scenario, etc.?).
- We need translators for some water quality standards (e.g., making narrative criteria numeric for attainment decisions).
- We need to know what action should be taken by a State if there is a violation of an aquatic life criterion (e.g., list under 303(d), report under 305(b), track elsewhere?).
- Concern about how fish advisories are considered in 303(d) listing; need to address this.
- Will guidance provide *de minimis* exceptions for Section 303(d) listing? (For example, for legacy problems/ atmospheric deposition)? *(Note: EPA staff indicated that this would probably be inappropriate but that actions to follow up on the listing could be tailored.)*
- CALM must identify specifically how to determine when a water body is non-attainment (e.g., what frequency of sampling is needed, what number of exceedances?).

How is existing and available information considered?

- Need to be clear about what "consider" means in context of 303(d).

- Non-regulatory factors should not be considered in assessments (e.g., ERL/ERM and sediments).
- Guidance should clarify how and when to use different categories of data (e.g., older data, inferences from land use info, reference to similar water bodies).
- Guidance is needed for waters that have non-traditional types of impairment (e.g., habitat problems); these should be addressed in a category separate from waters exceeding numeric pollutant criteria.
- Guidance should provide for considering effluent data from point sources in assessments.
- Guidance should emphasize monitored data, not rely too heavily on evaluated/modeled data.
- Modeling should not be a basis for Section 303(d) listing; impairments are generally outside range of validated/calibrated models.
- We are concerned about using biological data for Section 303(d) listing because there is no cause-and-effect relationship to enable the impairment to be resolved; TMDLs cannot be developed for biological impairments.
- To use biological data in making attainment determinations, there has to be a specific promulgated standard first.

How does a State define quality of data?

- Guidance should set time limits on use of different types of data.
- Guidance should address the need for rigor in scientific data.
- Guidance should provide QAPP (quality assurance plans) for dischargers, others to follow in collecting data.
- If data are not collected with quality assurance/quality control (QA/QC) in accordance with guidance, it should be discarded.

What data interpretation methods does a State use?

- It is important to assure transparency of impairment/attainment decisions so public can understand.
- Use the same process to decide impairment as to decide attainment.
- EPA should provide examples to States on data interpretation.
- Guidance should encourage States to reach out to stakeholders, including point sources to get the advantage of stakeholders' input on methods for defining impairment, conducting assessments.
- Guidance is needed on use of RF-3 and designation of segments (too many segments, not enough data coverage).
- States need qualified scientists to do statistical tests, interpret data and models.
- States should not rely on off-the-shelf models that use limited data.
- Guidance is needed re: interpreting/calibrating/validating models.

How does a state integrate multiple types of data in water quality assessments?

- Guidance should address the "pollutant" vs. "pollution" issue: how to integrate water quality data and assessments.
- We are concerned about EPA's policy of "independent application" of different types of water quality standards' provisions in making attainment determinations.

Comprehensive Assessment

How does the State ensure comprehensive monitoring coverage?

- Guidance needed on getting representative samples.
- States need sufficient data to assess water quality from both temporal and spatial perspectives.
- Requiring "comprehensive assessments" may raise concern that point sources will be required to increase their in stream monitoring; concerned about potential burdens.
- Would it be possible to allow a POTW that currently conducts extensive effluent water quality monitoring to switch to ambient water quality monitoring, assuming its effluent monitoring results show stable, compliant trend? Are States willing to 'back off', at least to some extent, on effluent monitoring by point sources to get more ambient data?

How does a State conduct ambient water quality monitoring?

- States should have flexibility to use alternative statistical approaches.
- We need guidance re: definition of water body segments.
- We are concerned about the frequency of data collection.
- There is a spatial bias to monitor near certain types of sources in current distribution of monitoring stations.

How does a state integrate additional data?

- We support use of locally collected data.
- We need guidance re: linking data from upstream/downstream reaches.
- Guidance is needed re: working across jurisdictions.

Data Management and Presentation

- It is important to educate public about what can be done, what to expect.
- We are concerned about the cost of building and maintaining comprehensive monitoring system(s).

Other

- We need guidance on analyzing impacts, social and economic cost/benefits under Section 305(b).

- We need specific guidance re: information required to delist waters under Section 303(d).
- The same test and rigor should be used to list/delist.
- The delisting process needs public input.
- We are concerned about the Section 305(b) "partially supporting" waters category; concern that this category will trigger Section 303(d) listing.
- Concern about how to handle "other" categories of impairment (e.g., legacy problems).
- Water quality standards need to be clear, well-understood. They should be numeric and measurable, not narrative.
- We need a document that gives technical basis for assessment of relative contributions of different pollutants.

Meeting 2 Notes:
Non-governmental Organization Representatives

General Comments

- Guidance should tell states how to monitor unmonitored waters and should clarify what states need to do. (Note: guidance is not binding on States.)
- CALM guidance should anticipate that biological and sediment criteria will be available shortly.

Attainment Decisions

Is standard attained/not attained?

- Guidance is needed on interpreting narrative criteria (e.g., to protect wetlands and aquatic life, anti-degradation) and determining what constitutes impairment.
- What value does probabilistic monitoring/sampling have in determining attainment?
- We need guidance re: what constitutes attainment of water quality standards (especially in light of limited data sets).
- Some evaluated data are more reliable than others and can show serious indication of impairment.
- Section 303(d) listing puts waters in spotlight both for restoring water quality and for revising water quality standards.
- How do you know if you have the data to make the right decision about revising water quality standards (e.g, to develop a site-specific water quality criteria)?
- As waters are identified as impaired, what is the right mix of actions to get the right criteria in place?
- Who is responsible for developing translators? Who will pay?

How is existing and available information considered?

- EPA should revisit its 1991 TMDL guidance that identified 19 categories of information.
- Narrative biological criteria must be a basis for determining attainment and listing under Section 303(d).
- Evaluated data should be considered and may be sufficient for listing a water body under Section 303(d).
- Older/aged data should be used until new data are available.
- If there is conflicting data (showing attainment/nonattainment), must be conservative; use data that indicates a problem in order to be adequately protective.
- Probabilistic monitoring and fish advisories-do you list under 303(d)? *EPA staff indicated that where sufficient data supports these findings, yes.*
- If a State puts a water on its Section 319 nonpoint sources list, possibly to get funding to restore it, but does not list that water under Section 303(d), there is a disconnect that must be addressed.

How does a State define quality of data?

- States should report about the quality of data that are used, whatever the quality is. The public has a right to know this.

What data interpretation methods does a State use?

- Attainment decisions need to be based on consistent interpretations (e.g., if 10% of samples exceed limit, water is impaired).

How does a State integrate multiple types of data in water quality assessments?

- The "independent application" approach should be used for Section 305(b), perhaps not for 303(d).

Comprehensive Assessment

How does the State ensure comprehensive monitoring coverage?

- Guidance re: requiring permittees to monitor ambient conditions would be useful.
- Environmental groups will support additional funding for water quality monitoring.

How does a State conduct ambient water quality monitoring?

- What size river segment is suitable for listing on the basis of evaluated data?

How does a State integrate additional data?

- State bias against using citizen-collected monitoring data must be overcome (many citizens follow State QA/QC procedures). NOTE: State monitoring councils may help.
- If a point source can demonstrate consistent compliance through effluent monitoring, we would support their undertaking ambient monitoring in exchange for less frequent effluent monitoring.

Data Management and Presentation

- Raw data used in developing a State's Section 303(d) list should be published with the list.
- States should clearly identify water body segments (geo-reference).
- States should be open re: what is on table, what we know, why we are/are not using these data; should have transparency in decision-making.
- Must identify waters States have not monitored.
- States should do a better job identifying what percentage of waters have/lack water quality data.

Other

- Could a single TMDL be developed to handle all waters in a state that are similarly impaired for the same designated use (e.g., air deposition of mercury causing fish advisories)?

Meeting 3 Notes:
Nonpoint Source Representatives

General Comments

- The nonpoint source community is interested in working on this effort.
- Currently, the 303(d) list of impaired waters is/should be rigorous and the 305(b) report is not as rigorous/quantitative in assessment. How will these lists be meshed?
- The nonpoint source community is nervous about getting on the Section 303(d) list because of the regulatory consequences (especially given that a water can be listed based on very little data).
- How will EPA handle outcomes of a CALM guidance that contradict current EPA policy, such as independent application?
- We appreciate that, through CALM, EPA is trying to raise the bar for the quality of data needed to list under Section 303(d).
- We like the idea of combining the 303(d) list and 305(b) report on a four-year cycle.

Attainment Decisions

Is standard attained/not?

- We need guidance on the amount of data that should be used to make decisions. (The "I don't know" list suggestion from the point source meeting is a good one.)

- Independent application is appropriate for 305(b), not 303(d) list.
- How should States make attainment decisions with limited data? Is one guilty until proven innocent or the reverse?
- Need to better define how to get on and off 303(d) list.

How is existing and available information considered?

- There should be a set of standard data that EPA requires/ accepts, with allowance for regional differences.
- How are ESA occurrences (habitat/range) used in CALM process? Are water bodies involved in ESA listings prioritized?
- Should aged data ever be used?
- Substances with MCLs are not always considered. This is a significant gap in the water program.
- States should use data from other sources and get over their bias against interest groups' data, as long as it is QA/QC'd. States should look at current information from, for example, land grant universities
- Biological information should be used to define designated uses and inform the standards process; look forward to seeing "Stressor identification evaluation" guidance (due out by end of 2000).
- Regarding monitored versus evaluated data, need to be clear on whether/how evaluated data should be used in attainment decisions?

How does state define quality of data?

- The regulated and regulators should come to agreement on what data are acceptable in each state.
- Better data will not necessarily mean fewer listings, although some people do not understand this.
- We need definition of credible data: scientifically valid biological, chemical, physical data collected with QA/QC procedures (includes historical data).
- EPA is concerned about defining "credible data" in such a way that it excludes otherwise useful data; see a need to be more flexible.
- Biological issues are important; need better bench-marking and scientific expertise on this; QA/QC procedures would help.
- Concern about 303(d) list being inaccurate; it must be scientifically valid.

What data interpretation methods does a state use?

- How do/should we assess non-monitored waters?
- How do/should we use data in statistical analyses?
- How do/should we use models to determine attainment?

- How do we distinguish/allocate impairments from flow or other stressor (physical, chemical)?
- How do/should States assess trends with limited data?

Comprehensive Assessment

- How can monitoring be designed to capture nuances in water bodies? (Need to use models to help with this.)
- How would we make sure the "I don't know list" is followed up on?
- Safe Drinking Water Act and Clean Water Act need to be more efficient/better integrated in doing evaluations.

Data Management and Presentation

- Use database approach to analyses: start general and drill down to specifics.
- The challenge is to communicate the issues to the non-expert public and also communicate what they (the public, including nonpoint sources) are supposed to do.
- We encourage EPA/States to make documented, transparent decisions.
- Guidance should define adequate "transparency" (better ways of communicating that the lists are out and what they mean).
- Proposed 303(d) list organization in New Jersey allows nuances to come through without putting all water bodies on 303(d) list. (This came out of a consensus process.) Possibly could be used as a model in CALM. Different categories in list:
    a. Have proof of impairment.
    b. Reason to suspect impairment, but not enough proof.
    c. Data contradicted intuition.
    d. Impairment, but not sure why.
    e. TMDL developed, but not fully implemented.

Other

- It is important to provide for dialogue between nonpoint sources and regulators without worry/threat of regulation. (Need to rely on more voluntary measures.)
- Voluntary BMPs can be used to improve water quality; with more money/support, many nonpoint sources will implement BMPs.
- Technology-based standards and guidance need to be updated and improved.
- Need to ensure that non-impaired waters stay clean (anti-degradation).

Water Quality Standards:

- Standards should be better defined to reflect designated uses (e.g., is it fair to use drinking water standards for aquatic habitat?)
- States should revisit water quality standards to make them more or less specific as appropriate.
- Standards should be more consistent across States.

- How will the CALM affect the triennial standards review? *[EPA staff indicated that as data come in, focused revisions to standards will occur and will be prioritized. Designated uses are the highest priority.]*
- Designated uses are important to improve, expand upon.

TMDLs:

- Section 303(d) list should go through second public comment period at the regional level.
- We should use 303(d)/305(b) processes to prioritize TMDLs.

Meeting 4 Notes:
State Agency Representatives

NOTE: Attached in Appendix C are:

1. *A written summary of their comments prepared by participants from the North Carolina Division of Water Quality; and*
2. *Issue papers submitted by representatives of the New Jersey Department of Environmental Protection, who were unable to attend the meeting.*

General

- We support the concept of consolidating 303(d) and 305(b) assessments.
- Consolidation makes sense from a workload standpoint; however, 303/305 are different tools/programs that maybe should not be merged. The 305(b) assessment is a snapshot of water quality, whereas 303(d) has regulatory implications and warrants more rigorous assessments. Section 303(d) and 305(b) need different sampling designs to answer different questions.
- Cross-state consistency is not paramount; need to be more concerned with internal consistency in each State.
- It will be a challenge to make interstate assessments consistent; the current time line may be inadequate to achieve desired consistency.
- We want consistent methodology, but need flexibility in criteria (e.g., DO).
- Consistent methodology across states help members of the public.
- We are concerned about inconsistency in reviews by EPA regional offices.
- Be cautious about how 303(d) and 305(b) are consolidated (lest the 303(d) list turns into list of impaired waters vs. waters requiring/needing TMDLS). Section 303(d) should not capture everything, but instead should be viewed as a subset of 305(b) "not fully supporting" waters.
- Section 305(b) more accurately reflects what we know about water quality.
- Use 305(b) assessment guidance as the baseline for assessments; rely on State/EPA 305(b) consistency workgroup work, too.
- Provide for both consistency and flexibility; make guidelines functional, not prescriptive.

- Re: guidance design, list questions for states to respond to, provide examples to set up boundaries, but allow states to give different answers.
- Statutory time lines will drive what we do.
- Timing concern: staff engaged in various related processes; there may be inconsistencies with other criteria/info (e.g., nutrients, SIE).
- Need to get input into CALM from other States, Tribes, River Basin Authorities.
- Need to involve EPA regional coordinators in CALM process.
- Need to involve other federal agencies in the CALM process.
- EPA regional staff should interact with state biologists to increase consistency at staff level through technology transfer.
- What we do creates feedback "do-loop": what happens in monitoring may affect water quality standards (and other programs).
- CALM may cause water quality criteria to be called into question in some states.

Attainment Decisions

Is standard attained/not?

- We agree with the need for transparent decision-making regarding impairments and for each use assessment type, but we are concerned about EPA micro-management of how decisions made, how much info used); ultimately, this should be at states' discretion.
- Suggestion: try to resolve guidance on attainment before addressing comprehensive assessment (as a first step, ask states to write up their current approaches).
- What is EPA's definition of "impairment"?
    - water quality standards-can violate x times without impairing use?
    - does D.O. due to salt wedges qualify as an impairment?
- What is the definition of "standards" for purposes of making attainment decisions?
    - designated beneficial uses;
    - anti-degradation provisions; and
    - criteria (numeric and narrative)?
- Need guidance re: differentiation between assessments and attainment decisions.
- EPA should extend maximum flexibility to states in interpreting narrative (and some numeric) standards; don't place extreme emphasis on numeric alone.
- We are concerned about defining impairment due to pathogens. Pathogens are indicators of sanitary conditions (Part II).
- Some States use monitoring data for 305(b) to determine use support and then use this information to develop 303(d) list. Would like to know if this a common approach across States?
- Many States rely on inferences to assess level of use support but this information cannot answer many questions (e.g., extent of impairment).
- Levels of rigor for listing and delisting should be equal; often, delisting has been subject to more intense scrutiny.

How is existing and available information considered?

- What qualifies as "existing and readily available info"? (Please provide clear examples.)
- Benthic macroinvertebrate information may not point to specific pollutant of concern.
- Don't expect states to agree on several issues (e.g., use of fish advisories). This can be dealt with via transparency.
- We need to encourage collection of habitat data.

How does state define quality of data?

- What "rules" will EPA use to judge quality? These should be written down in advance.
- Need to consider QA in whether data are used, but it may be tough to determine quality in some cases.
- Data quality, not the source of the data, should determine whether/how specific data are used in assessments.

What data interpretation methods does a state use?

- If EPA doesn't like the state's assessment methodology, will it reject the state's 303(d) list and promulgate its own list instead?
- We need guidance re: dealing with uncertainty.
- Oklahoma uses assessment support protocols (and we hope CALM is consistent with our approach).
- We are concerned about guidelines for listing and reporting prescribing different methods.
- "Pollution vs. pollutant": need guidance about assessments and the level of proof needed to differentiate "pollution" from "pollutant" for Section 303(d) listing decision purposes.
- Spatial distribution/applicability: how far can you extrapolate? What is considered representative?
    - When listing under Section 303(d), may be only guessing at the extent of impairment, but this has an impact on the size/geographic extent of the TMDL.

How does a state integrate multiple types of data in water quality assessments?

- Don't base attainment decisions strictly on numeric criteria; take a "weight of evidence" approach, consider best professional judgement.
- Clarify what is meant by "independent application" in the context of water quality standards.

Comprehensive Assessment

How does the state ensure comprehensive monitoring coverage?

- Resources for more, better monitoring are needed (at our agencies and elsewhere).
- CALM provides an opportunity for EPA to encourage monitoring and develop incentives to improve monitoring (e.g., by offering greater flexibility to more robust sampling programs).

- For states with extensive inaccessible wilderness areas, what does 100% monitored mean? [How can we use representativeness, extrapolation, probability-based approaches?]
- We need an answer as to what resolution the entire state must be monitored?
- 5 yr Basin-wide cycle offers a comprehensive look at basin and provides for public involvement; don't jeopardize this.
- It is cost-prohibitive to monitor every stream.
- Depending on design, rotating basin approaches could reflect condition of monitored waters, not condition of all waters of the state.

How does a state conduct ambient water quality monitoring?

- We need to show where are stream miles/types of activities and sources: need more targeted design.
- Guidance is especially needed on monitoring/assessing lakes and reservoirs.
- Section 305(b) guidelines are not designed for probabilistic monitoring. CALM should address how to use probabilistic sampling design.

How does a state integrate additional data?

- Monitoring councils that include the various agencies/stakeholders involved with monitoring allow state to reach out to stakeholders, may produce more and better data.
- Many states are already working with partners to integrate data but it takes a lot of time.

Data Management and Presentation

- We all struggle with how the state can allow for stakeholder input. Many coordination needs, many stakeholders. How can we do this best?
- We are concerned that States can never provide comprehensive monitoring coverage; need to tell the public this.
- EPA should not promote including all raw data in 303(d) lists; rather, provide a summary and make raw data available on request (contrary to the NGO comment).
- Need to avoid negative spin on water quality assessment documents (e.g., not an "Atlas of Polluted Waters"): positive presentation is helpful in keeping things moving forward.

- Interested in scale of WATERS database and flexibility re: how land uses can be combined; relieved that data can be entered into system at whatever scale is available and can be mapped at a finer level.
- EPA should communicate information at whatever scale states use (or, at a minimum, link to state website (UWA)); when EPA "rolls up" finer scale data to a national map, water quality often seems worse than is the case (especially for those states collecting data at a higher resolution).
- Interstate basins need to be addressed consistently in how data are communicated and what level of information/detail is provided to the public.

Other

- Section 303(d) list is a TMDL to-do list; concern that statute would override not developing TMDLs for parts 2, 3, and 4.
- De-listing should be iterative and ongoing.
- 'Sources and causes' information will come together in TMDL development (not through 305(b)).
- The iterative process to restoring water quality is most cost-effective: (1) determine limiting factors; (2) apply BMPs in sub-watersheds; (3) go back and monitor.
- There is a significant amount of money in proposed budget for 106 and 319: how should that money be spent?

305(b) Consistency Workgroup Meeting
Exploring the Connection Among WQS, 305(b) Reports and 303(d) Lists
October 17-19, 2000
Hyatt Regency Crystal City
2799 Jefferson Davis Highway, Arlington VA
(703) 418-1234

Note: this meeting will not address or discuss the new TMDL regulations.

MEETING APPROACH

Welcome to the October 2000 meeting of the Workgroup on 305(b) Consistency. Attached are the meeting **Agenda** and **List of Speakers.**

The meeting's **focus** will be on:

- **Monitoring for Comprehensive Assessment and for Listing Impaired Waters** (Tuesday afternoon); and
- **Assessing Water Quality Standards Attainment Status** (Wednesday).

Information on characterizing causes and sources of impairments will be presented briefly on Thursday morning.

On Thursday, participants will concentrate on the **issues of greatest concern** arising out of discussions on the two previous days. Three or four such issues will be identified by the participants for further deliberation in break-out sessions and a wrap-up plenary session.

In addition to facilitating an exchange of information, EPA's Assessment and Watershed Protection Division (AWPD) wishes to obtain **as much input as possible from each participant** on the meeting topics. Because approximately 120 State and EPA representatives are expected to participate in the meeting, plenary session discussion time will necessarily be limited. Therefore, participants are encouraged to take advantage of all of the following opportunities to express their views, communicate with EPA decision-makers, and share information with each other:

- **Participate in Open Facilitated Discussions following each panel presentation;**
  These discussions will allow not only for questioning of panel members, but also for brief suggestions, observations and information exchange among all participants on the relevant topic.
- **Complete and submit responses to the questionnaires in your packet;**
  Responses will be summarized and used to select issues for concentrated discussion during Thursday's break-out and plenary sessions. Responses will also receive further consideration by EPA after the meeting.

- **Participate in Break-Out Sessions on Thursday;**
  These sessions will explore key issues in more depth than would be possible in the full meeting. Break-out groups will be asked to develop options and recommendations on the key issues for EPA's consideration.
- **Ongoing Communication.**
  AWPD welcomes your continuing input after the meeting.

Ground Rules: The meeting will be as informal as the size of the group allows. When providing input, simply keep in mind the need to be brief, stay on topic, and respect differing views. Have fun!

Speaker List

October 17, 2000

Describing the Quality of States' Waters for 305(b) Reporting: To What Extent are State Waters Meeting WQS?

Tom Van Arsdall, KY—Progressing to a More Comprehensive Monitoring Coverage in Kentucky: Using Targeted and Random Networks in a Multi-Agency Watershed Approach

Michael Arcuri, WV—West Virginia's Approach to Monitoring and Assessment

Linda Schmidt, IN—IDEM ‘s Use of Probabilistic Monitoring Results for Comprehensive Watershed WQ Assessment: Advantages & Limitations of this Sample Design Program

Linked Monitoring to Target Waters

Dave Chestnut, SC—Identifying Impaired Waters as Part of a Monitoring Strategy

Richard Shertzer, PA—Pennsylvania's Statewide Waterbody Assessment Program: A Biological Approach Linked to GIS.

Nancy Immesberger, NJ—Integrating 305(b) Comprehensive Assessment Guidance with State 303(d) Responsibilities: Follow-up Monitoring and Listing Based on Extrapolated Assessments

October 18, 2000

What Metrics or Parameters are Appropriate for Assessing WQS Attainment Status?

Dave Chestnut, SC—Representative Parameter Suites/Indicators for Assessment of Specific Uses; Advantages and Disadvantages of Different Indicator Types for Assessing Specific Uses

Evan Hornig, USGS—Developing a Staged Approach to Parameter Selection for Assessing Aquatic Life Use Support
Neal Kammen, VT—Vermont's Use of Fish Tissue Data on Mercury to Assess Impacts to Fish Consumption Use
Documenting Data Quality

Bob Bukantis, MT—An Overview of Montana's Sufficient Credible Data Scoring Process

Diana Marsh, AZ—Designing an Approach for Implementing Arizona's Credible Data Law: Developing Data Quality Screening Levels; Identifying Level of Data Needed to Support Different Decisions
Jack Smith, WY—Wyoming Credible Data Statute: the Nature of the Legislation; Misconceptions; and How Wyoming DEQ will Implement the Provisions
Interpreting Data to Assess WQS Attainment Status: Statistical Tools

Charles Martin and Len Shabman, VA—Statistical Assessment Tools and Virginia DEQ's Application of the Binomial Approach

Daryll Joyner, FL—Florida's Approach to Interpreting Exceedances of Water Quality Criteria for 303(d) Listing Purposes

Interpreting Data to Assess WQS Attainment Status: Use Assessment Approaches

Derek Smithee, OK—Oklahoma's Use Support Assessment Protocol: Lessons Learned

Jason Heath, ORSANCO—Approach/Methodology for Public Water Supply Use Assessments
Wayne Davis, EPA—Key Data Interpretation Needs for Using Biological Data in Water Quality Standards Attainment Decisions: Lessons from MAIA
Doug Burnham, VT—Development and Implementation of Macroinvertebrate and Fish Community-Based Decisions: Assessing Aquatic Life Support Based on Deviation from the Reference Condition for Selected Wadeable Streams

Integrating Multiple Types of Data for Attainment Decisions

Jim Pendergast, EPA—Clarifying Misconceptions about the Independent Applicability Policy

Perri Phillips, MT—Integrating Data of Multiple Types and Quality to Make Use Support Determinations: An Overview of Montana's Process
Cynthia Grafe, ID—Idaho's Process for Integrating Multiple Types of Data to Make Use Support Determinations
Al Hindrichs, LA—A New Assessment Protocol for Dissolved Oxygen; the Need for Improved Coordination among Standards, Assessment and Implementation for More Effective Water Quality Management

October 19, 2000

Discussion of Strategies to Characterize Causes and Sources Contributing to Impairments

Sue Norton, EPA—Stressor Identification Protocol for Identifying the Causes (pollutants/stressors) Contributing to Impairments of Biological Communities in Aquatic Systems

Gregg Good, IL—Documenting Decisions for Attributing Cause and Source Categories to Impaired Waters

Facilitated Discussion

Summary of discussion following the panel "Describing the Quality of States' Waters for 305(b) Reporting: To What Extent Are State Waters Meeting WQS?"

Theme: Some states do not have the resources necessary to implement monitoring designs for comprehensive assessment.

- Resources vary significantly from state to state.
- The size of monitoring staff and the amount of funding are both key factors.
- The size of monitoring staff in the states presenting at this session ranges from 15 to 21 individuals. Other states noted that they have fewer staff.

Theme: Targeted versus probabilistic monitoring designs.

- A state representative raised the concern that probabilistic monitoring alone does not help a state to locate impaired waters that were not sampled. In response, EPA pointed out that no less coverage is provided than when a state uses only a targeted design. In that case, there are still impaired waters that do not get sampled.
- Some states remarked that their managements do not think probabilistic monitoring is necessary. The focus for states has been site-specific, primarily on TMDL development, which probabilistic monitoring does not address.
- States discussed the representativeness of targeted and probabilistic monitoring sites, such as how far results should be extrapolated from a sampling site. There is no clear answer. Some states are using more conservative estimates now than previously. The distance extrapolated should depend on land use and the size of the stream. One state pointed out a potential issue in that targeted sites are always selected for a certain habitat, but probabilistic sites are not always selected for habitat; therefore there may be a problem mixing the data or assigning results to entire reaches.

- One state asked whether probabilistic designs extrapolate survey results based on stream order. In other words, are results averaged and applied to the total number of stream miles in the watershed, or are the results for first-order streams applied to a greater percent of the total miles? If not, this is a weakness. EPA responded that in general the results are averaged and no weighting is generally given to stream order.
- States expressed concerns over the 303(d) implications of probabilistic monitoring. Some states do not want to have to do a TMDL for a site that was sampled as part of a probabilistic survey. Perhaps probabilistic data and sites not be made available to the public. EPA remarked that while Section 303(d) requires all data to be considered, it does not have to be used. Therefore, a state can explain that the data were taken for a different purpose and are not appropriate for 303(d) listing. There was some debate over whether doing a TMDL for a probabilistic site would bias results and prevent the site from being useful in future surveys.

Theme: Should probabilistic monitoring be used for uses other than aquatic life (e.g., recreation, fish consumption, drinking water, etc.)?

- At least one state uses probabilistic monitoring for fish consumption use. Other states indicated that probabilistic monitoring is used only for ALUS.

Theme: 303(d) listed waters should be consistent with 305(b) waters.

- There was a general consensus on this issue.
- EPA remarked that where probabilistic monitoring shows a greater percentage of impaired waters than targeted monitoring, this means the state needs to locate more impaired waters for 303(d) listing.

Facilitated Discussion

Summary of discussion following the panel "Linked Monitoring to Target Waters"

Theme: QA/QC of monitoring data.

- Some states are implementing annual training events and audits. EPA is seeking review of a draft document that addresses QA/QC of monitoring designs (See http://www.epa.gov/quality/qa_docs.html).

Theme: Involving ground water data and getting ground water folks on board.

- Some states are seeing a greater level of effort to get the ground water community involved, but are still not getting data that they can use.

Theme: Coastal and estuarine data.

- Issue for many states is how to describe the aerial extent associated with their off-shore monitoring. Guidance is needed.
- Many states report only beach or shellfishing (pathogen) data due to lack of resources.

Theme: Bacteria - "can we lay the issue to rest?"

- States remarked that a swimming closure is associated with a different level of pathogens than the water quality criterion for fecal coliform that many states have adopted into their water quality standards. Violations of the standard may not preclude swimming.
- One state commented that closures are often related to other causes, such as wildlife, pets, etc. Another state presented an example where dog waste is causing standard violations. What can states do in cases like this?
- Another state explained that because of the TMDL process, the public is demanding fecal typing. The state is working with the EPA Region to develop criteria for different levels of contact (e.g., immersion, beach, etc.).
- The issue of new indicators (fecal strep, enterococci) was raised. States expressed a need to have EPA support for making assessments based on different data. One state related that despite adopting new criteria, the problem hasn't gone away, only the sources have changed (e.g., dogs, horses).

Theme: Combining probabilistic and targeted monitoring

- Several states agreed that a combination of approaches is needed to address both the "big picture" and 303(d) issues. However, a number of states said they can only do targeted monitoring because of resource issues.

Theme: Data sharing

- States have sometimes not been able to use data shared by other agencies as much as desired due to different monitoring objectives, QA/QC requirements, etc.
- The public doesn't understand the amount of work required for a state to use 3rd-party data in 305(b).
- A number of attendees have had success in working with 3rd parties to meet the state's QA/QC and data needs. Some successes have included asking volunteer groups to change what and how they monitor, creating QA/QC coordinator positions, and meeting with stakeholders.
- Another approach is to use reports from 3rd party groups to guide state assessment decisions and future monitoring rather than subjecting the 3rd parties to rigorous QA/QC requirements.
- The federal government just published a unified federal policy that directs federal agencies to cooperate with states in the assessment of watersheds and requires consistency with state standards.

- Several states have developed monitoring councils to help coordinate and encourage data sharing and adequate QA/QC. EPA supports this approach and the work of the National Monitoring Council.

Facilitated Discussion

Summary of discussion following the panel "What Metrics or Parameters Are Appropriate for Assessing WQS Attainment Status?"

Theme: Need for approved methods.

- One state said it is a problem to wait for EPA approved methods in 40 CFR Part 136 before developing water quality standards, .e.g., for enterococci and E. coli.
- EPA's view is that states do not need an approved method to adopt a water quality standard. However, EPA is moving forward with methods for Part 136. Methods for enterococci and E. coli are out there. There are methods, but not through the Part 136 process.
- EPA is pushing hard to move forward with new indicators because of a new law (the "beach bill") which includes a requirement for coastal states to adopt latest EPA criteria by April 2004. There is a goal for states to adopt new indicators into standards by 2003.
- Regarding pathogen criteria, EPA is looking at full body contact. Shellfish folks say there is no epidemiological-based action level for enterococci and E. coli for shellfish consumption use. There are regional differences that need to be addressed. The Gulf Program will explore these issues as they pertain to shellfishing use.

Theme: Mercury TMDLs

- Why should states spend their limited resources on difficult-to-do Mercury TMDLs since it is a regional or national problem? Can EPA have national or regional TMDL approach for Mercury?
- EPA response: Mercury is one of the toughest situations. EPA is now looking at TMDLs that are being required and may develop an approach or guidance for doing this nationally. Not available now, but EPA is working on it.
- There are gradients of Mercury deposition across the country, and not all watersheds process Mercury the same way. Even if EPA can do large TMDLs, the state needs to be involved. A case study is the Savannah River in terms of helping states model using their own watershed characteristics.

Discussion Theme: Relationship between 305(b) and 303(d)

- Water quality assessors generally know when a water is impaired, but now due to the link between 305(b) and 303(d) listing, need rules spelled out, e.g., to "make the biologists lay it out on paper."

- A state expressed the opinion that there should be different thresholds for impairment under 303(d) and 305(b). Another state sees 305(b) has having multiple facets of reporting, but the use attainment component should be 1:1 with 303(d). That is, partial and nonsupport should equate to 303(d) listing, with the same criteria.
- EPA asked for feedback on the partial supporting category under 305(b).

Discussion Theme: The use of ambient toxicity testing data in 305(b) assessments or 303(d) listing

- Only a few states use ambient toxicity testing. Texas does list waters based on ambient toxicity testing; EPA Region 6 does the testing for them. Also, there is some limited use in Kentucky. Alabama uses it primarily to consider de-listing streams where benthics showed impairment. Connecticut did limited acute sediment toxicity analysis on highly impacted streams (samples analyzed by Region 1) and basically got no toxic response; they have not found it to be very useful because of insensitivity of test.

Discussion Theme: The use of threatened/endangered species data in 305(b) assessments or 303(d) listing

- In a New England hydroelectric project, researchers have measured high mercury levels in endangered loons. They use in weight of evidence approach that includes threatened and endangered species.
- Only one state mentioned using USFWS Endangered Species Act (ESA) listing in 303(d) listings and a few states mentioned considering ESA in weight of evidence evaluations.
- As EPA write guidance, should there be different criteria for threatened/endangered species versus non-threatened/endangered species? Response from one state was that criteria should be the same, should encompass all species; if endangered species are present, pay more attention, but do not change approach to assessment/listing.
- One state mentioned taking steps in WQS to develop management strategy for threatened and endangered species where there are water quality concerns. They will work with USFWS on this issue through basinwide planning.
- EPA noted that as a first guide, criteria to protect aquatic life uses should be used. If a state has info on how to protect a specific species, need to take that into account.
- One state had a Section 7 consultation with USFWS re: endangered mussels and learned that their metal criteria were not protective enough.

Discussion Theme: Partial/nonsupport and major/moderate/minor categories.

- A state expressed the view regarding partial support that public impression is important. Just "good and bad" is not adequate.

- EPA should decide whether major/moderate/minor categories for causes and sources of impairment are needed anymore. These categories were not used in the 1998 National Water Quality Inventory Report
- One state suggested that the tiered approach should work well, and asked what would happen regarding delisting if a state went that route, with prior listings made using old approach. Some waterbodies that were listed under the old approach would not be listed under a tiered approach.
- It is EPA's view that when the science supports new and better methods, they should be used, even though such changes may be a challenge to explain to the public.
- It would be hard to communicate a "maybe" category, and would need to be done carefully.
- This topic will be taken up in one of the break-out sessions.

305(b) Consistency Workgroup Meeting
Facilitated Discussion

Summary of discussion following the combined panels "Documenting Data Quality" and "Interpreting Data: Statistical Tools"

Theme: Credible data

- States are generally enthusiastic about establishing a credible data approach.
- One additional benefit of Montana's approach is having the information available to the public.
- Several states expressed a need for national consistency on the credible data issue so that all Regions are following the same guidelines with their states. Guidelines should deal with both monitored and evaluated data.

Theme: The binomial approach to making use support determinations.

- States using the binomial method do not correct for extreme conditions or for high-magnitude exceedances.
- VA does not use the binomial method for toxics; FL plans to do so and noted that their WQC are based on chronic criteria. Neither VA nor FL have used the binomial approach for delisting.
- In general, less than 20 samples is considered statistically weak for making use support determinations based on the 10% exceedance rule, with or without the binomial method.
- Several states expressed concern that waters should not be listed based on only 1 or 2 exceedances; metals are often the issue.
- Several states are interested in the binomial approach or other statistical approaches for dealing with small sample size.

Theme: Refining the 10% exceedance rule for non-toxicants.

- EPA's Water Quality Standards program would like to work on the issue of refining the 10% exceedance rule based on frequency, duration, and magnitude of exceedance.

Theme: Consistency between 303(d) and 305(b)

- Currently, several states automatically include 305(b) impaired waters on their 303(d) lists even if the assessments are based on minimal data (i.e., low confidence).
- Data poor states may tend to have inconsistencies between 305(b) impaired waters and 303(d) listed waters due to inadequate data to support listing decisions.
- Internal consistency within a state is sometimes an issue regarding what to list or not list. Guidance should create consistency.
- One state mentioned putting their 305(b) list out for public comment and another state mentioned involving the public in assessments via basinwide workshops.
- One option for a new category is "Fully Supporting, Impacts Observed."

305(b) Consistency Workgroup Meeting
Facilitated Discussion

Summary of discussion following the panel "Use Assessment Approaches"

Theme: Setting reference conditions.

- One member questioned the validity of eliminating 25% of reference sites in setting reference conditions. The danger is that later, when sampling unknown sites, the agency might find them impaired when they are not.
- EPA staff noted that in establishing reference conditions, the initial sites are always viewed as candidate reference sites. We don't know which sites truly make up reference condition. EPA sees this approach as building confidence and accuracy, not tossing out data.
- Western states reported problems with reference condition when reference sites have been devastated by forest fires. How to take into account natural variability? Several views were expressed: (1) data for such reference sites should be taken out of the reference metrics for samples taken around the time of an extreme event; (2) this is part of natural variability of reference condition; (3) we should not revise reference conditions after forest fires (which are often started by human activity); (4) it is more important to look at what percent of streams are taken out of reference conditions (versus particular streams); (5) it is important to study how natural systems recover over time.

- Recent studies suggest that there may be a 30-year oscillation in natural events, requiring twice that long to detect these natural changes.
- Once you have reference condition, the grey zone between impaired and unimpaired can be difficult to assess. One state has used a consensus approach to define the grey zone. The agency was glad to have some guidelines to follow for making determinations in the grey zone.

Theme: Biocriteria.

- Many of the state staff at the session said they use macroinvertebrate data to make use support decisions. Two of the states have biocriteria. Approximately a dozen of the states at the session are trying to develop biocriteria.
- Several attendees asked EPA to put more emphasis on helping states develop these criteria to back up use support decisions.
- EPA does want to see states adopt biocriteria into their standards for ALUS. EPA has not developed a rule to require biocriteria, although it was considered. EPA hopes that states, as they use the data more, see the benefit of having such standards. In the meantime, states can use biological assessments to implement narrative criteria and develop translator mechanisms to support development of TMDLs and NPDES permits, as appropriate.
- While EPA is not be forcing states to develop numeric biocriteria, the Agency has helped states further their capabilities. If it were forced externally or internally, states would be in a position to do this much more quickly now. Several states are also considering biocriteria for fish.

Theme: Bordering states.

- EPA encourages states to look at what is happening over the border or go to interstate commissions. Inconsistencies in assessments and TMDLs different across state borders is a concern. States should talk to EPA Regions for information.
- Some states share borders with states in other EPA Regions. It is difficult to share data across EPA Regions.

Theme: Drinking water and fish consumption.

- In 1996, at least 15 states did not assess drinking water use. The Clean Water Act requires that all uses be assessed. EPA urges states to coordinate their assessments under 305(b) and SDWA/source-water protection assessments.
- OST is drafting a memo on fish advisories and 305(b)/303(d) will provide some guidance that takes into account the peculiarities of those programs. It will point out that waterbody-specific data is more important than other data for listing decisions.

Theme: Nutrients and eutrophication.

- The decision-making process on eutrophication related issues is a major challenge (second to bacteria). For example, there is the issue of whether criteria should be based on chlorophyll a or species. Six or seven states indicated that they wrestle with nutrient/eutrophication water quality standards issues.

305(b) Consistency Workgroup Meeting
Facilitated Discussion

Summary of discussion following the panel "Integrating Multiple Types of Data for Attainment Decisions"

Theme: EPA's Independent Applicability (IA) policy

- EPA wants examples from states of conflicts in attainment status based on different types of data.
- A state responded that although use attainability analyses (UAAs) will solve problems of needed refinements to water quality standards, workload is a huge issue. For example, if a state has many streams that don't meet DO criteria, they would rather do biological assessments than do thousands of DO measurements. That is, let biological data trump water chemistry data.
- EPA is looking at making IA more flexible. In defense of chemical data, bioassessments tell us a lot about today; chemical conditions allow us to predict the future.
- Different criteria have different benefits. One attendee asked that EPA be clear that chemical data are not always better for predicting trends and that predicting trends is not the main purpose of assessments.

Theme: Dissolved oxygen issues

- Several other states besides Louisiana have issues with naturally occurring low DO.
- A state with an extensive site-specific standards program expressed the view that it is worth the trouble to establish sound criteria. Otherwise the regulated community will force the state agency to spend too much time defending its WQSs.

Theme: Diatom assessment methods

- Will diatom assessment methods survive the test of time? Idaho thinks so and noted that this community is used in Europe. The state is working with national experts and will publish a paper in the next year. Diatoms represent lower trophic data, and are good for integrating chemistry. Also, it is a quick, easy sampling method--just scrape slime and send it in to the laboratory.

305(b) Consistency Workgroup Meeting

Report Outs from Breakout Group Sessions

Establishing a "Maybe List"

Theresa Hodges, KS Department of Health & Environment, presented for the "Maybe List" breakout group. The group organized their presentation into six basic questions and associated options/recommendations.

(1) *Under what conditions can waters be placed in this category?*

The breakout group suggested that waters meeting any of the following criteria could be placed by States in this category:

- Insufficient data to determine attainment status;
- State has low confidence in the data used to list water (possibly, because it was collected by a third party, does not have sufficient QA/QC documentation);
- Conflict in data types;
- Marginal/borderline data (i.e., the water is not clearly supporting/nonsupporting);
- State lacks a numeric translator for the narrative water quality criterion;
- Underlying water quality standard is in flux;
- Data were collected during extreme events which may/not be indicative of typical conditions; or
- Listing/attainment status decisions that are based on evaluated (vs. monitored) data.

(2) *Where do you place these waters?*

The breakout group strongly recommended placing these waters in a separate category under 305(b). Some members felt this category could be a part of the "fully supporting" waters under 305(b); several others disagreed. The workgroup did not think States should be required to list these waters under 303(d) but wanted to leave that decision to States' discretion.

(3) *What should this category be named?*

The breakout group did not come to any agreement on what the category of waters should be called. The following suggestions were made, however, with a consideration of how the category's name may shape the public's perception of what actions will be taken on behalf of the waters.

- Waters of Concern;
- Not Rated;
- Inconclusive;
- In Need of Verification; and
- Indeterminate.

(4) *Should the category of waters be published? If so, where?*

All breakout session participants agreed that this list should be published along with the 305(b) report. It should be left to States to decide whether to include these waters on the 303(d) list.

(5) *What followup actions are necessary for these waters?*

The group agreed that follow-up actions must be designed to address the factor(s) that initially triggered listing under this category in the first place. The following types of follow-up actions may be generally appropriate:

- Increase water quality monitoring (to deal with data quality/quantity issues, conflicting data, and waters listed based on data collected during/following extreme events);
- Verify water's compliance record and/or presence of other management plans;
- Review underlying water quality standard-possibly in context of triennial reviews (to address translator issues, standards in flux, and methods questions); and
- Develop a narrative criterion translator.

In addition, the group suggested tying in follow-up actions to existing basin management plans, monitoring strategies, or "active management lists" the State may already have. Finally, the group recommended that States be required to specify both the type and timing of their next action but not necessarily the final action related to the water in question.

(6) *What action should be taken by States that currently include these waters on their 303(d) lists but may not wish to do so in the future?*

The group recommended that EPA allow States to remove these waters from the 303(d) list if they meet any of the conditions or criteria listed under Question 1, above.

NOTE: This group also briefly discussed how this category of waters would impact State listing of threatened waters. Ultimately, the group believed that "breaking out" these waters would tighten and clarify the threatened waters list(s).

Data Quality/Statistical Tools

Al Hindrichs, LA Department of Environmental Quality, presented for the Data Quality/Statistical Tools breakout group. As background, Mr. Hindrichs reported that fundamentally, EPA "approves" State methods through it's approval of the 303(d) list and Section 106 grants. He then reported that the group identified the following priority need.

EPA should develop guidance on making water quality attainment decisions, especially for small, chemical data sets collected during routine ambient monitoring.

The workgroup recommended that the guidance:

- Build upon existing guidance (e.g., the data quality/data quantity hierarchy in the 305(b) guidelines);
- Be tailored to address specific parameters (e.g., metals, pathogens);
- Be based on how criteria were set and uses applied; [NOTE: Certain criteria (e.g., human health) may be more "important" and therefore warrant special attention or consideration.]
- Identify different/appropriate statistical methods for different sample sizes;
- Recommend that States consider their confidence in datasets/analyses when making attainment decisions [NOTE: "Important" criteria need stringent confidence requirements."]
- Consider the variability of the parameter used to determine attainment status and the representativeness (e.g., seasonality, age) of the data in monitoring design and analysis.

The breakout group also recommended that EPA do the following:

- Evaluate the merits and limitations of the various statistical options available;
- Evaluate whether data and/or analyses with different confidence levels can be used to support different decisions (e.g., supporting, not supporting, may support...);
- Clarify whether alternative approaches (other than the "10% rule") will be allowed; and
- Make development of the guidance and the analyses described, above, a high priority for the Agency

Monitoring Design

Jay Sauber, NC Department of Environment and Natural Resources, led the presentation summarizing the breakout group's discussion. Mr. Sauber opened his presentation with the following observations and comments made by the breakout group.

- More resources are needed.
- There are differences in the data quality needed for and purpose of 303(d) lists and 305(b) reports.

- Monitoring designs must be evaluated within the context of the primary goal of the effort. For example, States generally have very specific, targeted data and reporting needs (related only to the State); EPA often takes a more holistic or 'big-picture' approach, using the data to report on waters of the nation or to develop Reports to Congress.
- States tend to favor targeted monitoring programs; EPA's water quality monitoring program data needs are best served by probabilistic monitoring programs.
- EPA should not divert new Section 106 (and other) funds to develop TMDLs.
- Probabilistic monitoring data are difficult to use to support site-specific listings under 303(d) or 305(b).
- Many good data sources exist. Important sources of data include:
    - volunteer monitoring information;
    - State data;
    - EPA's Environmental Monitoring and Assessment Program (EMAP);
    - NPDES discharge permit reports;
    - U.S. Geological Survey;
    - Fish and Wildlife agencies (State and federal);
    - Nongovernmental organizations, including conservation groups, volunteer groups;
    - Universities;
    - Natural Resources Conservation Service;
    - U.S. Forest Service; and
    - National Oceanic and Atmospheric Administration.

- While good data exist, environmental agencies (at the State level and at EPA) lack the staff to pull together the information at the State or national level.

Mr. Sauber then noted that, historically, 303(d) lists and 305(b) reports were based, to varying degrees, on speculative information (especially related to identifying causes and sources). The onset of TMDL program lawsuits, however, has increased the degree of scientific rigor needed to justify a 303(d) listing. States have a higher confidence in their data and in the statistical tools they use to interpret the data.

Next, Mr. Sauber stressed that it is important that *303(d) and 305(b) are not inconsistent*. This is critical even though 303(d) and 305(b) "ask" different questions (level of impairment vs. impact as compared to a reference stream) and require different levels of "defensibility."

The group focused much of its discussion on the following questions:

For 305(b): Do you have a comprehensive assessment of your waters?

The breakout group determined that "comprehensive assessment" does not necessarily mean that each and every segment of water has been assessed but rather that a State has a comprehensive understanding for the different waterbody types (e.g., lakes, rivers,

estuaries). To answer the question, the breakout group suggested, a State should be able to do the following:

- Describe what it means by "comprehensive";
- Explain how it 'covered' the State (e.g., percent of waters monitored, evaluated, assessed);
- Describe its monitoring program (what constitutes a good assessment vs. an excellent assessment). Good monitoring programs may include the following elements:
    - multi-matrix assessments (biological, chemical, etc.);
    - adequate spatial and temporal coverage;
    - targeted monitoring;
    - rotational basins; and
    - seasonal coverage.

    These elements should be described for all relevant waterbody types, including rivers, lakes, streams, estuaries, and beaches (although the breakout group was unsure how to do this for wetlands).

    Excellent monitoring programs may include the following additional elements:

    - probability-based monitoring to cover all waters;
    - long-term, statewide coverage;
    - coverage of all waterbody types; and
    - information on maintaining and continuously improving the collection of data and information.

For the 303(d) list: Can you tell where impaired waters are and why they are impaired?

The breakout group suggested that, to answer this question, a State should be able to provide the following information:

- timeframe;
- confidence levels (e.g., 90%);
- extent of impaired reach (extrapolation up and downstream); and
- an explanation of how the State made the use support decision.

hrid-Party Data

This breakout group focused on the use of third party data to inform or support 303(d) and 305(b) decisions. Jack Smith, WY Department of Environmental Quality, presented for the breakout group. Mr. Smith reported that the breakout discussed the following key questions:

How does a State use its third party data?

Who solicits these data?

Which third party datasets are good enough to use to support 303(d) listing?

The breakout group developed the following recommendations related to third party data use by the States.

- States should invite all third party data..
- States should identify in writing how they plan to use the data (and for what purpose).
- States should require different QA levels for the different data uses. For example, data used to support 303(d) listing decisions should pass stricter QA/QC tests than those used to support the development of a State's 305(b) report.
- States should clearly state their QA/QC requirements for data.
- States should lay out their formatting preferences (e.g., electronic) for data delivery.
- States should clearly state their preferred level(s) of analysis (i.e., should the third party submit raw data or data that have been analyzed in some manner).
- States should advise the public and other interests well in advance of when the data are requested.
- States should communicate their plans for using third party data through established state channels, such as:
    - watershed councils;
    - land management agency partners;
    - special interest ties; and
    - public notice.
- States should follow through on their written policies.

Summary of Breakout Discussion Session

Following these presentations, the moderator (Martha Prothro) opened a general discussion by asking participants to reflect on how their managers would respond to these recommendations/possible new directions presented by the four breakout groups. Several participants responded that their managers would be pleased (1) that these issues had been discussed, (2) that many of the recommendations called for flexibility, and (3) that EPA was planning to offer a variety approaches (good to best). At least one participant suggested that State program managers would be more comfortable if EPA call this document a "technical resource" rather than "guidance."

One participant raised the issue of whether public citizens (and, especially, special interest groups) might be concerned that the "maybe" or "indeterminate" category would turn into a "dumping ground." Several meeting participants commented that they already

set aside certain waters and have found public involvement in and understanding of the listing process helps allay others' fears. Others noted that it is important also to be clear about what followup activities are planned for the "indeterminate" list waters. One participant emphasized the importance of convincing constituents that, with this category, the state is able to focus attention on a broader universe of waters than it otherwise would. At least one state representative noted that environmental groups in his state already do not trust his agency and would rather see the waters on the 303(d) list.

One meeting participant voiced a concern about the emphasis of statistical tools, and wondered how his State would fit the biological and habitat information it collects into the statistical approaches discussed by the Data Quality/Statistical Tools breakout group. This person commented that "raising the bar" for statistical certainty would make it difficult, in his state, to show impairment. Another participant suggested that EPA (or some party) prepare and maintain a resource guide of experts on multi-metric indices, statistical tools, and other topics.

Several participants noted that volunteers often collect high quality data, and observed that these data, if they have undergone a QA review, are often appropriate to include in 305(b) reports. At least one participant suggested that States build "front-end" systems to accept third party data.

Summary of December 2000 Stakeholder Meeting on the Consolidated Assessment and Listing Methodology (CALM) Initiative

Overview

This document summarizes four meetings hosted by the U.S. Environmental Protection Agency (EPA) on December 4 and 5, 2000, in Chicago, Illinois, at the EPA Region 5 Offices. These meetings were a follow-up to meetings held in June 2000 in Washington, D.C. to obtain stakeholder input on the CALM initiative and generally followed the same format as the June meetings. Subsequent to the June meetings, EPA's Assessment and Watershed Protection Division (AWPD) published a draft outline of Phase I of the proposed Consolidated Assessment and Listing Methodology (CALM). Participants in the meetings were asked to comment on the draft outline.

The CALM initiative addresses identification of impaired waters under Section 303(d) and preparation of water quality assessment reports under Section 305(b) of the Clean Water Act. Phase I would focus on water quality standards attainment decisions and comprehensive quality monitoring programs.

In this document, general themes emerging from the meeting are identified first, followed by separate summaries of each of the four meetings.

The meetings were designed to obtain input from stakeholders as follows:

Meeting 1: Point Source Representatives
Meeting 2: Non-Governmental Organizations
Meeting 3: Nonpoint Source Representatives
Meeting 4: State Agency Representatives

At each meeting, each participant was given an opportunity to speak so that EPA could consider all points of view. This document therefore captures the diverse comments of individual participants and cannot be construed as representing a consensus of the participants or any group of them.

Participants' observations were recorded on flipcharts during the meetings. During and at the end of each meeting, participants were asked to review the flipcharts for accuracy and completeness. With minor changes to improve clarity for those who were not present, the detailed summaries of the separate meetings contain the original flipchart language. Participants' comments are organized under key subject headings and are not necessarily presented in the order they were offered at the meeting. During an introductory session at the outset of each meeting, Margarete Heber, Chief of the Monitoring Branch in AWPD, briefed participants on the draft outline. [A copy of the slides Ms. Heber used during her presentation is included at the end of this summary as Appendix A.]

Agenda

Each meeting followed the same basic agenda:

(1) Introduction of participants and review of meeting objectives;
(2) Brief presentation on the CALM initiative and draft outline by Margarete Heber;
(3) Facilitated discussion between participants and EPA staff, focusing on participants' views;
(4) Opportunity for comment by non-participant observers; and
(5) Wrap-up/review of next steps.

The meetings were facilitated by Ross & Associates Environmental Consulting, Ltd, with notetaking support from Tetra Tech, Inc. Each participant was specifically invited to speak during the meeting and almost all did so. The meetings were facilitated to ensure fair opportunity for all to speak, to help keep to the schedule, and to help stay on subject. However, most of the discussion was informal. From time to time, EPA staff answered questions or asked questions of the participants.

Participation

Participation was generally limited to representatives of the interests for which each meeting was organized (point source representatives at Meeting 1, non-governmental organization representatives at Meeting 2, etc.). Participants generally decided independently which of the four meetings to attend. In a few cases, participants in a meeting may have represented interests other than those targeted for that session.

Participants were seated at a table, along with EPA representatives. Each meeting was open to the public and several observers were present at all meetings. The number of participants from the targeted stakeholder group for each meeting was as follows:

Meeting 1: 31 Point Source Representatives;
Meeting 2: 6 Non-Governmental Organization Representatives;
Meeting 3: 10 Nonpoint Source Representatives; and
Meeting 4: 16 State Agency Representatives.

Over the course of the two days of meetings, 78 individuals attended, including EPA staff, observers, and meeting facilitators. Attendees are listed in Appendix B of this document.

Themes Emerging from the Four Meetings

General Reaction to EPA's Plans to Develop Guidance on a Consolidated Approach

- Most participants expressed **support** for the overall approach. (This and most other comments were consistent with comments raised at the June meeting.)
- Several participants noted that the draft outline was not specific in acknowledging the **differing purposes of Sections 303(d) and 305(b)** of the Act that may require different approaches to attainment determinations (e.g., amount and quality of

data needs may differ). Some participants expressed concern, often based on resource limitations, about adopting different monitoring approaches for 305(b) (probabilistic monitoring) and 303(d) (targeted monitoring). Others agreed that such an integrated approach would be reasonable.

- Some participants favored including **benchmarks** in the CALM document to help stakeholders determine if a state has a strong monitoring and assessment methodology as compared to other states. State participants, however, generally strongly favored flexibility in this regard. Some participants recommended incorporating a regularly scheduled **evaluation** into the CALM effort to ensure that states are following the methodologies they submitted to EPA.
- Many participants commented on or asked questions about the degree of **specificity and prescriptiveness** in the CALM document. Some (mostly point source representatives) suggested that EPA be specific about: 1) minimum standards for acceptable state programs; and 2) what is not acceptable. These concerns focused primarily on Section 303(d) listing decisions. With regard to comprehensive monitoring programs, a few participants pointed out that minimum standards would assist with state budgeting and resource allocation decisions. Others raised concerns, however, that if EPA defines minimum elements of a state monitoring program, some states may fund only these activities, possibly reducing current investments in monitoring.
- Some participants noted that the purpose statement in the draft outline seems to focus on the need to improve documentation. While participants agreed that this is an important purpose, they suggested that a more important and overriding purpose is to **improve data quality** to ensure better understanding of water quality and better decisions about attainment status.
- Many participants acknowledged that the CALM effort, if successful, will help **foster an increased level of trust** among EPA, states, the regulated community, and the public by providing for an **open and transparent process** for attainment decisions. Some state participants expressed concern about possibly inconsistent and/or overly harsh review and evaluation of their decision- making processes by EPA Regional Offices. One suggestion was that EPA Headquarters should issue guidance to the Regions on how to review responses from the states on their methodologies.

<u>Determining Attainment/Nonattainment</u>

- Consistent with suggestions made at the June meeting, various participants supported the CALM's document providing information on:

  - How to handle **historical/aged data**;
  - General data requirements and process for **303(d) delisting**;
  - Proper uses of **evaluated data** generally;
  - When/whether to use **biological data** to determine attainment status;
  - The definition, intent, and use of **EPA's policy of "independent applicability"** when integrating multiple data sets to determine attainment status.

- In addition, participants suggested that EPA address:

  -How to make decisions about nutrient and (clean) sediment problems while waiting for **nutrient and sediment criteria** from EPA;
  - How to determine attainment of designated uses and narrative criteria;
  - How to address wetlands criteria and/or antidegradation provisions. (It was noted that these may be addressed later than in Phase I.)

- Many participants noted that the process of making attainment/non-attainment decisions is predicated on **scientifically-sound water quality standards**. Many participants urged EPA to work to improve the process for review and development of water quality standards. Several participants expressed concern about making 303(d) listing decisions based on narrative criteria.
- Most participants said that states should establish quality assurance/quality control (QA/QC) parameters that could be followed by stakeholders willing to perform monitoring and that any data, including **stakeholder data**, should be used without bias by states in making attainment determinations, as long as the data are collected in accord with the prescribed parameters. However, many state participants expressed concern about the cost to states of using third party data. Participants suggested that it might be helpful to route third party data through a watershed group or other NGO with data quality assurance mechanisms in place.
- Many participants noted that it is important for state and EPA attainment determinations to be as **transparent to the public** as possible. A transparent process will foster **locally-led solutions** to many water quality problems. Participants also noted that EPA and the states should better define the purposes and consequences of Section 303(d) lists and Section 305(b) inventories to the public.
- Many participants requested guidance on **303(d) delisting** decisions. Participants representing point and nonpoint source interests specifically suggested that EPA provide for states to revise their current lists to reflect the data quality/quantity and decision processes states establish in response to the CALM.

<u>Comprehensive Assessments</u>

- There was **general support** for the concept of ensuring that all waters are assessed and there was considerable interest in the probabilistic monitoring approach for Section 305(b) assessment purposes. Participants encouraged EPA to consider state monitoring as a priority for **Section 106 funding**.
- Some participants suggested that, through the CALM process, EPA should provide states with incentives to **perform more monitoring**.
- Participants requested clarification on the "nested monitoring" approach described by EPA that encourages the use of **probabilistic monitoring** followed by **targeted monitoring**. Many participants encouraged EPA to clearly define each monitoring approach and describe how and when states should use each approach. Participants generally suggested that probabilistic monitoring approaches are

appropriate only for use in 305(b) efforts and that states should use a more targeted approach for making 303(d) listing decisions.
- Several participants from the point source community expressed an interest in **voluntarily conducting ambient monitoring** to supplement state efforts. However, some participants from this group expressed concerns about states requiring ambient monitoring in NPDES permits.

For more specific information on views expressed by participants, please read on to the separate meeting notes, below.

Meeting 1 Notes:
Point Source Representatives

General Comments

- CALM is a worthwhile effort. We are in need of a clear, transparent, scientifically-based methodology in each state. EPA giving guidance makes sense and is important. We might quarrel about specifics, but we agree it needs to be done.
- CALM is worthwhile, but the methodology for 305(b) and 303(d) will need to be different. We may need two documents. If so, a document for 303(d) is priority.
- The tone of the outline does not specifically match the comments and emerging themes of the last meeting. For example, the outline should include more discussion on the fact that 303(d) should be based on more statistically sound, site-specific data and on numeric water quality standards. Also, the outline should specifically include the idea that 305(b) indicates where we need to collect more data. Waters should only be listed where there is hard data and numeric criteria. The outline currently contains some "squishy" statements regarding 303(d) listings (e.g., listings not based on numeric water quality standards are ok). This draft outline seems more "data negative" and concedes the fact that we don't have enough valid data.
- The point source community pays attention to 305(b) and sees value in 305(b) to identify problem areas. We view the 303(d) list as a subset of the 305(b) report.
- POTWs are willing to be involved in ambient monitoring as a tradeoff for effluent monitoring.
- POTWs are willing to monitor water quality as a way to protect the public's investment in effective wastewater treatment.
- Point sources are opposed to increased requirements for permittees to collect ambient data. Prefer voluntary approach to participating in ambient monitoring.
- The states should make water quality standards, implementation guidance, and the schedules for upcoming triennial reviews available online. EPA promised this, but it has not yet been done. The assessment/attainment determination process begins with the water quality standards.
- The document resulting from the outline should not be referred to as a "guidance document" because that implies that states do not have to follow it. It is not clear

that there is an incentive for states to change if these concepts are not included in regulation.
- Placing state water quality standards on the web is relatively easy and will likely be done in time, but interpretation of the standards is difficult to explain on the web. That is why processes and rationale should be documented as CALM would provide.
- Point sources are most interested in 303(d) because of its implications to permittees. EPA should link 305(b) and 303(d), then the 303(d) listing would derive from 305(b). EPA should emphasize that there are different consequences associated with each of the lists. The document should clearly state that decisions made on the basis of data related to 305(b) and 303(d) have different regulatory implications. Not all lists are equal.
- This document should not focus only on better documentation of decision processes; we want incentives for increased monitoring.

Attainment Decisions

- Include a discussion of delisting with the discussion of attainment in the CALM document.
- The June CALM presentation had a strong emphasis on measured data. The outline did not have the same emphasis. The outline should emphasize the importance of using measured data to make decisions.
- Concerned about the "maybe list"concept. States would need this to be part of their 305(b) reports. This is new ground. Previously there was no room for inconclusive data or "maybe." Waters were either attaining or impaired.
- Clarify that only numeric standards and hard data can lead to 303(d) listing.
- Not enough emphasis on measured data. Using actual hard data is the right way to make listing decisions.
- Clarify Section 13.4 of the outline which lists 3 ways of determining extent of attainment. Are these options, and if so, is there a preferred option? (EPA is seeking documentation on what states are doing now and plans to provide guidance on what is acceptable in the future.)
- Will the concept of a "maybe list" be used? (EPA stated that 303(d) does not authorize a "maybe list," but in their 305(b) reports states may identify waters that lack sufficient data to make an attainment decision. These waters would be candidates for follow up monitoring.)
- Clarify whether/how threatened waters would be addressed.
- What if a water is on a past list, but the state's new methodology says the data are inconclusive? Would the water go off the list and be placed in another category for follow-up monitoring? (EPA stated that under the current regulations waters may be removed from the list only if the original basis for listing was mistaken or when the state determines that the listed water is in attainment. Third party data, such as from citizen volunteers and point sources, could help to document attainment. Some states will put these waters at the top of their 303(d) list because they think when they actually get data, they'll be able to delist rather than develop a TMDL.)

- If a water is listed as a result of only one data point, it should not stay on the list.
- Section 2.2 of the outline includes mention of documentation of data quality as an important piece of reporting, but does not include details on including back-up documentation. Helpful to include a list of "metadata" that is necessary to document data quality, as in **(Section 3.3?)** of the outline.
- Section 3.4 of the outline addresses statistical tools for interpreting data. Concerned that states may erroneously list water based on small data sets. Waters should not be listed unless there are real data showing non-attainment.
- Regarding Section 9 of the outline, provide guidance on interpreting each type of data when integrating multiple types of data.
- Regarding EPA's policy of independent applicability, what if data on copper show an exceedance, but biological data says water is healthy? (EPA explained its independent applicability policy and stated that there is no known case where equally valid data for both chemical and biological data suggested different attainment decisions.)
- Clarify the quality and quantity of data in relation to application of the policy of independent applicability. The statements contained in the outline regarding the use of biological data seem to be in conflict.
- Clarify the minimum number of observations needed for an attainment determination.

Comprehensive Assessment

- Clarify the concept of probabilistic monitoring. EPA should be extremely careful to clarify the uses of each monitoring approach described in the outline -- probabilistic monitoring is for use in developing the 305(b) list and targeted monitoring is used for developing the 303(d) list.
- Include discussion of a basinwide approach in the monitoring design portion of the methodology.
- Some states are putting monitoring requirements in NPDES permits that go beyond compliance monitoring (e.g., WET, pollutant characterization, sediment, in-stream, etc.). Permittees would be willing to voluntarily conduct this type of monitoring and share data, but we would prefer not to have these monitoring requirements included in permits.
- Will EPA set minimum elements for state monitoring programs and provide opportunity to comment on that document?
- POTWs should not be required to collect ambient data, but the states should be open to including third party data and, when practical, include third party data in the design of the state's monitoring program.

Other

- Would like to see a web site that summarizes each state's assessment methodology, including the parameters, frequency, decision methods, etc.
- Explain "SIE" in the flow chart entitled, "Using Multiple Types of Data to Assess Attainment." (EPA explained that SIE refers to the Stressor Identification

Evaluation Guidance, a document that addresses identifying the cause of biological impairment. This guidance is due out soon. EPA agreed that all acronyms need to be defined.)

- We would like to know how many TMDL efforts have resulted in delisting instead of a TMDL.
- Will EPA involve the U.S. Fish and Wildlife Service (FWS) in the development of the document? (EPA stated that the FWS is involved in other programs, such as the TMDL program, but is not directly involved in the development of this document.)

Meeting 2 Notes:
Non-governmental Organization Representatives

General Comments

- Existing 305(b) guidance seemed perfectly clear - 10% exceedance means impairment. Is the existing guidance not good enough for 303(d) purposes?
- What are the consequences for a state not having an approved methodology?
- Is the point of CALM to encourage more monitoring?
- We support putting a state's methodology on paper for people to review; tends to move efforts in a positive direction.
- Like the question and answer format of document. Difficult to determine from the outline the number of questions that EPA will use. Recommend that EPA use less questions rather than more.
- The outline only mentions antidegradation once. Unable to find anything that addresses waters that are meeting their designated uses and water quality standards, but are not meeting antidegradation. (EPA stated that antidegradation will not be covered in Phase I of the CALM.)
- EPA should have a provision for states to assess how well they are following their methodology, in particular how they are using volunteer data. Recommend requesting this information from states every two years.
- Encourage NGOs to get involved in standards process to ensure states have good, complete standards. Everything is predicated on a state having good, accurate water quality standards.

Attainment Decisions

- Are rules for water impairment for 303(d) list based only on aquatic life uses? At least one state seems to be using this approach to shorten the 303(d) list. (Example: a water is impaired for fish consumption, drinking and swimming, but the state doesn't list the water unless the aquatic life use is not met.)
- Will CALM address naturally occurring background that is causing an exceedance of the water quality standards? Would a state list waters impacted by background? (EPA stated that this is dependent on how the water quality standards are written, and that states should look at this before going through the next listing process.)

- Concerned that states with credible data laws may not meet the 303(d) requirement to consider all existing and available data.
- Concerned that arbitrarily placing age limits on data, in an effort to use higher quality data, could create a severe disincentive for monitoring. In some states, waters are taken off the 305(b) list based on the fact existing data are old, although there are no new data and no clear reason for assuming the old data are no longer representative.
- EPA should address issue of treatability in Section 9.5 of the outline. Public water supply example where intake water meets the ambient criteria, but through treatment more total dissolved solids (TDSs) are added to the water, causing it to exceed the drinking water maximum contaminant levels (MCLs). As a result, the water does not support the designated use because it is not drinkable.

Comprehensive Assessment

Why is EPA asking states to define "waters of the state?" Has there been a problem with the definition of this term in the past? (EPA stated that states have historically focused on rivers and streams, and some of them exclude other navigable waters such as lakes, coastal waters, and wetlands from their definitions.)

- Observation: At least one state would like to use TMDL models to identify most effective monitoring stations.

What is the federal standard for acceptable third party data? (EPA stated that there is no federal standard regarding acceptable third party data.)

- Inclusion of citizen monitoring data is a primary concern. Is there no guidance for states to include these data? (EPA stated that CALM will include guidance on the use of third party data and that EPA has previously issued guidance encouraging the use of citizen monitoring data.)
- Source water assessments under the Safe Drinking Water Act produce data that could benefit 305(b) and 303(d).
- Water quality monitoring councils can bring together a multitude of federal and state agencies. It is important to clearly and narrowly specify objectives of a monitoring council and then broaden out over time.
- Some states are years behind in assessing data that they have collected, even though they have expanded their monitoring programs. At least one state won't take water purveyor data even though it has gone through rigorous QA/QC.
- Which states do not accept third party data? (EPA mentioned that some states are passing "credible data laws" and other laws that prohibit the state from using any third party data, including data collected by other state and federal agencies.)
- Use NGOs, such as river authorities, to get volunteer data through to the state. In Texas, there is only a team of 11 people to do assessments. Although volunteer monitoring data are available, allocating resources to conduct QA/QC of the data is a problem.

Other

- How do people get a sense of where their state ranks in comparison with other states to determine which states are doing a good job and which states are not? Is there a compilation of methodologies from all states that allow people to compare?
- Does this document take you through both monitoring and assessment?
- Would like guidance on how to test for mercury criteria.
- Will EPA address assessment of effectiveness of TMDLs? Will states shift the focus of monitoring from assessment (i.e., monitoring healthy waters that have not yet been monitored to ensure they stay healthy) to tracking implementation and performance of TMDLs and delisting waters on the 303(d) list?
- It is important that additional Section 106 resources do not get used for other activities besides monitoring.
- Clarify the portion of the outline that discusses metrics/indicators. Unclear that metrics and indicators are synonymous and that EPA is asking the state to identify indicators for each use.
- To what extent is the United States Geological Survey (USGS) participating? (EPA explained that USGS is part of EPA/state workgroups, particularly on monitoring design. Participating in development of a national probability-based monitoring design.)

Meeting 3 Notes:
Nonpoint Source Representatives

General Comments

- CALM was needed five years ago. It is an important effort. Concerned that numerous existing water quality programs seem to operate in a vacuum. Also concerned that many impaired listings are based on drive-by assessments.
- CALM provides states with guidance how to develop a more transparent process. Real opportunity to tell states that old data should not result in a listing that may not be needed. States are hiring contractors to collect data to confirm impairment/attainment and if a TMDL is necessary. A lot of money is spent on correcting TMDL listing problems that could have been spent elsewhere on priority environmental problems.
- CALM is a good idea. It is very much needed. The devil is in the details, and it could potentially be very expensive. Concerned about taking into account variability in waters, such as seasonality, when determining attainment.
- We support CALM but it should have been done five years ago before the 1998 list was developed. It is difficult to backtrack in a regulatory scheme. Now that we have the huge list, how do we deal with perception that we're ignoring the old list as we revise the methodology and reevaluate?
- Initially excited about the CALM process, but disappointed that purpose of the document, as stated in the outline, does not specifically include improving the

data upon which these decisions are made. (EPA agreed that this should be clarified.)

- Providing for documentation does not go far enough. Where are the benchmarks for determining where we need to be in making a program better? Will technical documents referenced in the outline provide these benchmarks? Provide list of references for documents cited in the outline so stakeholders can obtain these documents and review them. (EPA agreed that this should be done.)
- Do not use acronyms in the outline or be sure to include a list of acronyms. (EPA agreed.)
- Define "iterative process."
- Without the "meat" in the outline, it is difficult to tell if the CALM document is heading in the right direction.
- Locally-led processes for addressing water quality problems are a critical component of watershed management. Through the CALM document, EPA has the opportunity to create a transparent process to help local folks understand the process. Voluntary approaches should be used wherever possible.
- Does the phrase, "Need to achieve a robust assessment of water quality within resources available for monitoring" in Section 12.3 of the outline mean "robust as possible with the money you have?" The goal should be to put more resources into monitoring so it will be a robust program.
- In the State of Oklahoma, the CALM process has had a positive effect on making transparent decisions in several programs (e.g., 305, 303, 314, 319).

Attainment Decisions

- Will the CALM document address the refinement of water quality standards, specifically designated uses? This would help with the listing process.
- EPA Regions have required some states to use the 305(b) process to collect newer information on waters that lack information as a way to expand the 303(d) list. Using the 305(b) process to expand the 303(d) list can breed distrust.
- Be more precise in the suggestions on seasonality and age of data in the guidance. Otherwise, more state legislatures will enact credible data laws. (EPA responded that in some situations, conditions in the area have not changed and older data may still be relevant. States will have to decide on a case-by-case basis.)
- The State of Iowa just passed a credible data law that states any data more than 5 years old cannot be used for regulatory purposes. (EPA stated that states are required to "consider" all data; the state does not have to "use" them. It was also noted that the Iowa law provides a rebuttable presumption that older data cannot be used, so in some appropriate cases it may be possible to show that the data are still representative of water quality conditions.)
- Provide the states with guidance on what data should be used as historical data and what data are credible. It is difficult for folks to understand that we are making decisions based on data that may be over 10 years old. Older data can be used for certain purposes, but they should not be considered credible unless they are proven credible. The CALM document should include an approach that

addresses credible data/historical data issues. If this issue is not addressed, the document will not have gone far enough and state legislatures will step in.

Comprehensive Assessment

- Regarding the concept of "nested monitoring," is EPA saying that states are not required to monitor all waters, but will make inferences to determine impairments for other waters? Will states list waters based on information collected somewhere else? (EPA explained the purposes and appropriate uses of probabilistic and targeted monitoring. Probabilistic monitoring design gives statewide information on water quality and helps to determine where to conduct follow-up monitoring. This type of monitoring provides data for use in 305(b). Follow-up monitoring is targeted and helps states to make decisions related to 303(d).)
- EPA should be very explicit about the differences and purposes of probabilistic and targeted monitoring in the CALM document.
- How is the issue of seasonality addressed in probabilistic and targeted monitoring design? EPA should keep in mind different time scales relevant to impacts (e.g., daily to years). The timeline for impacts can be long term, for example silviculture harvests occur once every 25 years, in the interim the forest is just standing undisturbed.
- How many states are using probabilistic monitoring designs? What cycle are states using under this design? (EPA responded about 6 to 12 states use this type of monitoring design. Different states are using different cycles for achieving comprehensive assessment. EPA would like some detailed examples of the different ways states may set up probability design, timeframes for implementing them and options and limitations for aggregating these results over time.)
- Is there buy-in at high levels of EPA regarding CALM? Will some of the new Section 106 funds go to states for monitoring or will other programs take a portion of these funds?

Other

- Why is EPA pushing states to go forward with TMDLs when this effort is telling states to re-evaluate and strengthen the underlying monitoring programs?
- How does CALM play into TMDLs now under development using older data? States are using data collected in the mid 1990s to develop TMDLs. Will TMDLs developed now change in light of new data?
- Does an EPA Region need to have a methodology in place if it second guesses the list developed by a state and modifies the list? (EPA stated that the Region will document its decision.)
- What does EPA have in mind regarding the implementation of the nutrient and sediment criteria?
- How does CALM affect the current lists and TMDLs? This will have significant implications for land owners.

- How does an interested party provide comments on the federal document and on the methodologies submitted by the states to EPA. (To comment on the outline, go to the CALM website. The comment process on methodologies developed by the state will differ. Most states will have a stakeholder process as they develop their methodologies.)
- What is the QA 9 Document referred to under Appendix B of the outline?
- U.S. Department of Agriculture (USDA) is not involved in water quality monitoring since it does not have the charge for monitoring like USGS. USDA is concerned about BMPs and relies on USGS water quality data to determine BMP effectiveness. Due to the Government Performance and Results Act (GPRA), USDA is making an effort to record data on a watershed basis. Data are available on a county-by-county basis for 16 key issues/BMPs. The web site with this information is www.nrcs.usda.gov/prms. USDA wants to encourage a locally-led, voluntary approach that includes technical and financial assistance rather than regulatory approach under NPDES or TMDL programs. This is key to solving water quality problems.
- Local landowners are affected by listings, but do not always receive information on what to do or the cause of impairment. It is frustrating from a landowner's perspective. People need information. The states are so busy doing paperwork that they can't do monitoring. The watershed approach is supposed to be voluntary, but stakeholders are not being asked to volunteer. We want to do something. Those of us who could most impact the outcome are not involved.
- Is the ultimate goal to understand criteria and standards?
- Why don't we work on the 305(b) assessments and prevent waters from being listed? Address trend analysis. It is important to catch waters with declining quality.
- Strongly encourage EPA to address the issue of delisting and the issue of transition from the old way of doing business and the new CALMs that states develop.
- Does the definition of "waters of the state" include wetlands? (EPA stated that the definition does include wetlands and that EPA is working on developing monitoring protocols for wetlands.)

Meeting 4 Notes:
State Agency Representatives

General

- Extremely supportive of the CALM process. Concerned, however, that EPA is asking the states how they make decisions and calling it "guidance." This is not guidance. This is information gathering to begin a process. EPA is really asking the states to be transparent, which makes states feel vulnerable. Concerned that EPA Regions may review their methodology and may "slap" them for their process.

- Concerned that if EPA defines adequate elements of a state monitoring program, some states will view this as "minimum elements" and figure it is fine to fund only those elements.
- Include a definition of monitoring in the document. Some think monitoring means just sampling. The definition of monitoring should include everything from design to collection to analysis to reporting.
- In support of the CALM effort. It will help to iron out the consistency issues identified in New England states. Would like the document to cover both 305(b) and 303(d), so that impairment under one list equals impairment under the other list. Also, explore eliminating the difference between "partially supporting" and "not supporting" due to the problems caused by this distinction.
- CALM is a good idea since the same resources are often allocated for both programs. Concerned, however, that this could be too prescriptive.
- Asking questions of the states is good. States have not been asked questions in 10-15 years. It is good to revisit these questions and actually write down the answers. It is difficult for states to answer these questions, however, without minimum requirements.
- CALM is giving the states the opportunity to reinvent their programs.
- Good job EPA! This is the first document that identifies data analysis as a task or function.
- If this is a guidance document, EPA should make recommendations on what is good and what is acceptable. EPA should not say "must" - this term is too inflexible and tries to force fit programs.
- Identify in the document the advantages of monitoring at the state level and at the federal level (e.g., data should drive decisions of where to spend resources).
- Appreciate the pressure that EPA is under to assess waters under 305(b) without tipping the 303(d) listing.
- Asking for maximum flexibility through CALM.
- Appreciate EPA's efforts to be very direct, clear, and transparent. Please continue this. Concerned that this openness might erode down the road. Continue to be clear where EPA (specifically Margarete) has authority and where it does not.
- We have begun a process of building trust between EPA and states. Trust is at the root of all concerns. EPA has said that documenting methodologies is not a test, but we are concerned that it may be a test down the road. If there is a test to pass, we would like the answers to the test ahead of time.
- Thanks to EPA for its efforts.
- Sharing information on states' activities is valuable.
- Describe how things in Phase 2 of CALM will relate to transition issues and issues addressed under Phase 1 of CALM.
- Why does the presentation include very few details on Phase 2 of CALM?
- Need information in the document on how to handle issues during transition. For example, how do we make decisions until EPA publishes nutrient criteria.
- Encourage EPA Headquarters to issue guidance to the Regions on how to review responses from the states.
- It is important to move this process to Phase 2.

- Concerned about the language. It sounds like an integrated review of program elements. Would hate to see the minimum standards become the best of standards.
- Provide the states with minimum elements of a monitoring program.
- Be clear about the elements that should be included in the methodology.
- Clearly state the purpose of CALM in the introduction of the document.
- Important to emphasize that CALM includes lakes as well as streams, wetlands, and coastal waters.
- Include as many definitions and purpose statements as possible to help the states interpret the intent of the document.
- Confusion regarding the content of Part A and Part B of the outline. Revisit the issues that are part of assessment and part of attainment decisions.
- Part A of outline should focus on monitoring and Part B of outline should focus on attainment. (Reverse the current order.)
- Use questions to prompt states to think about how to deal with interstate waters that do not have interstate organizations to address them. Need to promote data sharing and attainment decision making.

Attainment Decisions

- We need to look at when water quality standards should not apply.
- Provide more guidance on delisting.
- Delisting guidance is needed. EPA has to address interim process with new regulations.

  Refinement of designated uses has a bearing on attainment/non-attainment decisions. Mention this issue in the methodology.

  Northeastern states are comfortable with the idea of a "maybe list."

- Is the idea of threatened waterbodies no longer in discussion?
- Although a "maybe list" has no legal basis, the state can create a "maybe list" based on how it writes its methodology.
- Include information on different methods of data analysis for 303(d), 305(b), and ambient background in CALM document.

Comprehensive Assessment

- Probabilistic sampling does not make sense for purposes of 303(d) listing. Are we still only listing segments or listing subwatersheds if sampling is dense enough?
- EPA is convening a workgroup to determine how to incorporate probabilistic monitoring into 305(b) reporting.
- Monitoring programs cannot consist only of probabilistic monitoring.
- Incentives to conduct probabilistic monitoring are not apparent.
- Isn't the goal to tell Congress that we are monitoring all waters?

- Continue targeted monitoring. Use biological monitoring to determine if there is an impairment and use targeted monitoring to determine the source of the impairment.
- Point sources are collecting a lot of data and feel badly when states tell them there is not enough money to analyze the data.
- Texas publishes a data submittal document so folks collecting third party data know data requirements ahead of time. Encourage other states to make data requirements available.
- Address third party data issues such as how to use it, when to use it, what to do with it. Address third party data in the context of resource concerns and QA/QC concerns.
- Encourage the use of third party data, but realize that incorporating these data is a very labor intensive activity.
- How can this be used to draw in other programs?
- Who is going to look at all available data? To do so is extremely burdensome.
- This is a good opportunity to bring in Section 319 and get funding for 319 lakes assessment.
- The outline does not discuss database requirements or needs. A system for storing and statistically manipulating data will advance the cause. There needs to be consistency in such a database.
- Will EPA require the use of new STORET under CALM?
- We need to do a better job of defining the lists to the public.

Other

- Is the purpose of this document to give guidance on how to develop listing methodologies?
- The State of Indiana has received comments that support and request more prescriptive methodologies.
- EPA Region 6 has said that Section 122 water quality standards that could affect NPDES must be reviewed and approved. Region 6 wants to approve states' listing methodologies.
- Will this guidance replace the existing 305(b) guidance or is it just laying out how states do business? Want to know what states need to do at a minimum; this information will help with budgeting and resource allocation.
- What is the timeframe on criteria development? What is the timeframe on BEACHES?
- Clarify the phrase "power of datasets" from the slide on Guiding Principles Part A.
- Would like a compilation of state methodologies and monitoring programs to understand what other states are doing.