

Part B lite QA/QC Review Checklist for Aquatic Vital Sign Monitoring Protocols and SOPs

Roy Irwin, NPS, WRD

March 4, 2008

Suggested citation: Irwin, R.J. 2008. Draft Part B lite QA/QC Review Checklist for Aquatic Vital Sign Monitoring Protocols and SOPs, National Park Service, Water Resources Division. Fort Collins, Colorado, distributed on Internet only at http://www.nature.nps.gov/water/Vital_Signs_Guidance/Guidance_Documents/PartBLite.pdf

Draft Revisions in Progress. These topics are complex and there will be additional updates and improvements in the future. However, the text may be used as a working draft. Please send peer review comments to roy_irwin@nps.gov

Table of Contents (In Word, a control-click on the chapter headings below will take you there):

Disclaimer:	4
Introduction.....	5
QUALITY ASSURANCE (QA):	7
I. Summary of Information from Past Data	7
II. Document Objectives and Questions	8
III. Document Vital Signs and Measures and How They Were Chosen	9
Always Measure Required Parameters	14
Which Nutrient Measures to Monitor:	18
IV. Include Detailed SOPS for All Field and Lab Methods	19
Consider Existing Protocols and Guidance from Other Monitoring Agencies:	19
Consult NEMI on Lab Methods.....	24
Put the Details in SOPs and their Appendices	25
Secondary Data Collection from Existing Data.....	26
V. Monitoring Design Summary in Protocol Narrative.....	27
Representativeness	28
Target Populations versus Sampled Populations	29
Refine Monitoring Design and Representativeness Iteratively	30
Consider Interagency Design Recommendations:	31
Representativeness versus Diel Water Column Measures:.....	33
Representativeness versus Tidal Cycle Signals:	36
Representativeness in Wadeable Streams:	36
If All Sites Were Selected With a Judgmental Approach.....	38
Causation.....	40
Stratification.....	40
GRTS and Similar Approaches for Assessing Status	41

Will a Probabilistic Monitoring Design be used for Status or Trends?	43
Introduction to Probabilistic Designs for Long Term Trends.....	44
GPRA and Proportions	47
Will the Information be Useful to Management?	47
Does It Still Make Sense?.....	49
After Revisions, Go Back and Optimize Related Sections.....	50
QUALITY CONTROL (QC):.....	51
Why Document Quality Control?	51
Include a QA/QC SOP and Comparison Table for QC Topics	55
VI. Completeness, Sample Sizes, Statistics, and Detection Probabilities vs. Desired Conditions	59
1) Refine (Provide More Time and Space Detail) Objectives and Questions	62
2) Identify Desired Conditions Qualitatively First	62
Moving From Qualitative To Quantitative Goals.....	65
Consider NPS Impairment Guidance.....	65
Consider O/E Goals	66
Iterative Goal Setting	68
3) Identify Resource-Collapse and Other Thresholds of Concern.....	69
4) Identify Existing Conditions.....	70
5) Develop Safety Margin between Existing Conditions and Threshold Magnitude ...	70
6) Document Variability in Time and Space	71
7) Revisit and Refine Target Population Details	74
8) How Big of a Difference or Change Do We Need to Be Able to Detect?	74
Monitoring Design Sensitivity vs. Measurement Sensitivity	74
Calculate Monitoring Design Sensitivity.....	76
Effect Sizes (ESs) Based On Multiples of the Standard Deviation	77
Minimum Detectable Differences (MDDs) in Original Units	78
9) What Initial Statistics Will Be Used?.....	81
10) Choose Desired Confidence/Detection Probability (Power = 1-beta).....	82
11) Choose Significance Level (alpha).....	84
12) Use Simple Calculators to Make Initial Estimates of Required Sample Sizes.....	85
Perform Different Initial Simple Calculations Depending on the Scenario:	85
Sample Size Calculations for Nonparametric Procedures	89
Before a Good Estimate of the Standard Deviation is Obtained:	90
After a Good Estimate of the Standard Deviation is Obtained:.....	91
Sample Size Needed to Detect a Defined Difference between Two Means.....	91
Sample Sizes Needed for Differences between Two Means When Using Paired Sampling	93
Sample Sizes for Two Samples, Variances Unequal	95
To Detect a Stated Difference between a Mean and a Standard.....	96
Solving for Minimum Detectable Difference Rather than Sample Size:.....	98
Inequivalence (Bio-inequivalence) Sample Size Calculations	98
Comparing Inequivalence, Equivalence, and NHST Options:	100
Sample Size Needed to Estimate a Single Proportion	102
Sample Sizes to Estimate a DIFFERENCE between Two Proportions.....	104
Cases Where Variances Are Not Calculated in the Usual Way.....	104

Composite Samples, a Special Case	106
Bacteria Sampling: Another Special Case	107
Transects, Another Special Case:	107
Sample Sizes Needed for Confidence Intervals in General:	108
Sample Sizes for Two-Sided Parametric Confidence Intervals about a Single Mean	110
Sample Sizes Needed for Confidence Intervals around DIFFERENCES between Means	113
Nonparametric Confidence Intervals about a Single Median	114
Sample Sizes and Statistics for Taxonomic Richness.....	115
Sample Sizes Needed for Trend Analyses	115
Rethink Detectable Difference Goals for Trends.....	119
13) When In Doubt, Throw It Out:	121
Don't Just Report an Unsatisfactory Result, Change Something:	122
14) Optimize Monitoring Plan Details for Affordability and Logic.....	122
15) Draft Initial Sample Sizes and Optimized Monitoring Design.....	123
16) Finalize Sample Sizes and Design with an Applied Environmental Statistician .	123
17) Estimate the % of Samples That Will Fail	124
18) Increase the Planned Sample Sizes Accordingly.....	124
19) Include Completeness Goals in a Table in the QA/QC SOP.....	124
VII. Data Comparability (Internal/NPS and External/Other Regional Data)	125
Comparability in Agreement or Pass/Fail Scores	126
VIII. Measurement Sensitivity	127
Low Level Detection Limits (MDLs and MLs).....	129
MDL:.....	129
Minimum Level of Quantitation (ML)	131
How Will Values below the MDL or ML be Reported and Analyzed?	133
Alternative Measurement Sensitivity (AMS) and AMS+.....	134
Difference between AMS and NIST Expanded Uncertainty	136
Difference between AMS and MDL.....	136
Difference between AMS and AMS+.....	137
Difference between AMS and Precision.....	137
Do We Need both AMS and MDLs?.....	137
AMS and AMS+ Reporting in STORET	138
Censoring AMS Values	139
AMS in Biology and Habitat Observations	139
How Often Should AMS or AMS+ be Calculated?.....	139
AMS Tools:.....	140
Resolution	141
IX. Measurement Precision.....	142
Repeatability or Reproducibility Precision?	143
Express Precision Results in These Ways:	143
Put Precision Details in a QC Table	144
Specify Precision MQOs as data Acceptance Criteria.....	144
Precision+	145
Sample Sizes Needed for Precision Estimates:.....	146

Split Sample Options to Estimate Precision	147
Precision Compared To Sensitivity and Detection Limits.....	147
X. Measurement Systematic Error/Bias/Percent Recovery	148
Include Calibration Details	150
Blank Control Bias (usually applicable to chemical lab work only)	151
NON-QC SOPS RELATED TO QA/QC	152
XI. Include a Data Analysis SOP.....	152
Confidence Intervals about a Single Mean	152
Confidence Intervals about the Difference between Two Means	153
Upper Confidence Interval Limit on a Single Mean.....	153
Confidence Intervals around Differences between Medians	155
Bacteria and pH Statistics, a Special Cases	155
Short or Long Term Trend Analyses?	156
Consider Diel Differences in Trend Analyses	156
Stratify or Weight by Flow or Water Level Before Trend Analyses?	156
Consider Phenology Factors in Trend Analyses	157
Consider Seasonal Differences in Trend Analyses.....	158
Too Many Choices for Trend Analyses?	158
Pseudoreplication Issues	159
Keep it Simple with Time Period Tests for Step Trends?	160
Regressions in Trend Analyses:.....	160
Missing Values, Useful Data, and Effective Data	161
Useful References for Statistical Analyses:.....	162
XII. Include a Cumulative Measurement Bias SOP	163
XIII. Include STORET Details in a Data Management SOP.....	168

Disclaimer: Nothing in the discussion below should imply government endorsement (or lack thereof) of any specific products. Commonly known products (including books now often used in our field) are mentioned strictly as examples, but there probably others out there that may be superior in various ways now or in the future.

The 4-Letter Acronyms Herein unless otherwise defined tend to be NPS acronyms for Park Service Units. A list of all parks and their 4 letter park specific acronyms in each NPS Vital Sign Monitoring network is available at NPS 2005. [Vital Sign Monitoring Networks](#). A list of [network coordinators and contact information](#) is also available. Since Network acronyms are harder to find on the internet, all network names are now written out herein.

Introduction

The discussion herein covers many of the same topics covered in the broader and less up to date version of [Part B](#) [Irwin, R.J. (2004). Vital Signs Long-Term Aquatic Monitoring Projects: Part B. Planning Process Steps: Issues to consider and then document in a monitoring plan including monitoring protocols and standard operating procedures (SOPs) for Quality Assurance/Quality Control (QA/QC). Water Resources Division].

Some found the original [Part B](#) to be too long, while others complained when it was shortened, suggesting explanations be included. Part B is still available as a broader resource that is aimed more at the whole planning process.

Part B lite (herein) is intended to be complete guidance **optimized for use when developing protocol narratives and attached SOPs**. Although an original goal was to make the lite version ‘short’, by popular demand sections that more completely explain complex issues have gradually been added. When the choice was between short vs. clear or complete, short did not prevail. This was done many times throughout the document and therefore, “lite” no longer means short.

Other NPS [WRD guidance documents](#) (Parts A, C, D. and E) are listed as links from NPS [WRD Guidance Documents](#).

As suggested in generic VS guidance (K.L. Oakley, L.P. Thomas, and S.G. Fancy, 2003. [Guidelines for long-term monitoring Protocols](#). Wildlife Society Bulletin 31(4), all protocols should include:

- A. Protocol Narrative
- B. Protocol Standard Operating Procedures (SOPs), and
- C. Protocol Supplementary Materials

An important item on any protocol review is whether or not the protocol follows the organization above, is complete, and has a table of contents that helps one determine where things are. We recommend that a QA/QC SOP be included that covers most of the topics covered herein. For those networks who want to follow state and EPA conventions, the QA/QC SOP could also be called a quality assurance project plan (QAPP) SOP.

Many of the QA topics covered in the first sections herein are touched on briefly in the protocol narrative, often with more detail in the QA/QC SOP. The [QC](#) topics (from comparability on down herein) can be covered primarily in the QA/QC SOP. Many of these topics are interrelated and the different pieces need to make sense when considered as a whole. Therefore, we recommend liberal use of “point-to” links to help readers understand the big picture and where the important pieces are.

Either the protocol narrative or a separate SOP should include a discussion of who will do the monitoring and who will train them and how often (recurrent training and is Quality Assurance/QA basic). Is there a SOP that clearly defines protocol variables and how to measure them?

The following text summarizes the basics of what has to be in water quality and other aquatic protocol SOPs to meet checklist ([Checklist for Review of Vital Signs Monitoring Plans](#), hereafter referred to as “the checklist”) requirements.

This summary can also be used for the basics that should be included in Phase 1, 2, and 3 monitoring plan chapters. In most cases, the planning process is iterative, with very general statements in the plan chapters becoming more detailed in the subsequent protocol narrative, and then even more detailed in SOPs.

Among the basics that need to be covered in the narrative and SOPs are the following monitoring basics (adapted from [USGS Managers' Monitoring Guide](#)):

WHAT are you going to measure?

WHERE are you going to put your sampling points?

HOW are you going to measure it?

WHEN (and how frequently) are you going to measure it?

One of the main originators of the survey design and response design concepts popular in EMAP and other survey design disciplines helpfully clarified the terminology distinctions as follows (Scott Urquhart, Department of Statistics, CSU, Personal Communication, 2005):

What: Sampled Population and/or Target Population

Where: Monitoring, Survey, or Sampling Design

How (and Who) -- The Response Design. The response design incorporates numerous decisions about how to measure the attribute of interest accurately (Larsen, D. P., T. K. Kincaid, S. E. Jacobs and N. S. Urquhart (2001). Designs for evaluating local and regional scale trends. *Bioscience* 51:1069-1078).

When (and how often) -- The Temporal Design (Although Larsen et al. 2001. op. cit. clarify that this should simply be part of the sampling design, others seem too often to simply overlook [diel](#) variation, changes with flow, and other important temporal aspects or aspects that directly or indirectly tend to drive variation or magnitude changes, often related to some temporal detail or upon some factor other than variability of changes over space. Even Larsen et.al. 2001 (op.cit.) seem to be stressing mostly looking only at changes in variability within one year versus across years, although they do state that “if concordant variation is high, neither revisiting sites within years nor adding sites can have much effect” (on getting smaller confidence intervals on trend magnitudes). .

Splitting the design into response design vs. sampling design components is relatively new in the literature. However, regardless of the terminology used, in modern scientific thinking (as well as modern environmental monitoring planning), quality assurance is now correctly recognized as not just something that one thinks of at the last minute, as an afterthought at the end of planning. Instead it is now more broadly understood as a process which should influence the entire planning process. This includes summarizing what is already known, carefully thinking through the specific questions that need to be answered and then making sure the new data to be collected are optimally

relevant, representative, comparable, and of adequate quality and quantity to meet study objectives.

On one level, QA/QC is simply a very methodical system of making sure the monitoring design and SOPs are defensible and “make sense.” See [Part B](#) (the longer version of this document), for additional more detailed discussions of the entire planning process, and each part, and how all the pieces fit together.

QUALITY ASSURANCE (QA):

There are various ways to define QA. In a typical example, EPA defines QA as “an integrated system of management activities involving planning, implementation, documentation, assessment, reporting, and quality improvement to ensure”...quality (EPA. 2006. [Guidance for the data quality objectives process](#). EPA/240/B-06/001.).

Adequate training and qualifications help insure adequate QA, but not by themselves. Most quality assurance components are not measurable and are thus qualitative rather than quantitative, whereas most quality control (QC) measurement quality objectives are measurable and quantitative. In other words, the control in QC is based on Performance-Based Measurement Systems (PBMS).

The International Union for Pure and Applied Chemistry (IUPAC) helpfully points out that “[Quality assurance](#) is meant to protect against failures of quality control.” In other words, if QA is not good enough, [QC](#) measurement quality objective standards may not be met. QA is therefore the guarantee that the quality of a product (analytical data set, etc.) is actually what is claimed on the basis of the [quality control](#) (QC) applied in creating that product. [QC](#) is then basically how one assures that the product meets or exceeds some minimum standard based on known, testable PBMS criteria.

Another key concept in modern scientific thought is that QA relates to all the qualitative things that are done to ensure quality in the whole systematic planning and project management process and is not just a last minute task one does at the end of planning. It includes carefully thinking through the questions that need to be answered after summarizing what is already known, making sure the data collected are relevant, representative, comparable, and of adequate quality and quantity, and making sure the study design is defensible and “makes sense.” All of the steps outlined below are part of QA, but only measurable performance characteristics for data quality indicators like measurement [precision](#), measurement [bias](#), measurement [sensitivity](#), and (for chemical measures only) [blank](#) control are typically also considered QC.

I. Summary of Information from Past Data

QA [checklist](#) question: For water quality monitoring, has information content of available past aquatic data (for each waterbody being considered for monitoring) been adequately summarized in terms of hints of trends or other important issues of concern? Networks should summarize available data including the data in NPS Horizon’s Reports ([Baseline Water Quality Data Inventory & Analysis Reports](#)). The word “hint” is used carefully here since old data is seldom perfectly definitive, complete, or perfectly comparable between agencies or time periods.

The emphasis should not only on what groups have been monitoring where and when, but also on “what does the data collected mean?” Again, what is the information content of the past data regarding hints of trends or issues of concern? Although this may have been briefly mentioned in chapter 1 of the central monitoring plan, typically more detail should be provided in protocol narratives.

Other than Horizon’s reports for parks, insight on hints of issues or trends based on past data can be found in various State, Federal, and Regional Summary Reports. For example, those looking at coastal waters can look at summary reports from NOAA ([National Estuarine Eutrophication Assessment](#) --NEEA monitoring program) and EPA (Marine EMAP [National Coastal Condition Reports](#)).

A table listing 303d waters should be included in the protocol narrative, along with a note that the most recent WRD [Designated Use and Impairments Database](#) (intranet link works on NPS computers only) has been consulted and that any differences with the vital sign network versions of the 303d lists have been logically reconciled. When possible, there should be more spatial detail (impaired from where to where?) in protocol narratives compared to related discussions in the background section in chapter 1 of the central monitoring plan of each Vital Signs Network.

II. Document Objectives and Questions

Most NPS VS monitoring networks are using objectives based generally upon the five generic [Specific, Measurable, Monitoring Objectives Goals of Vital Signs Monitoring](#) (status and trends in selected indicators, etc.). In water quality, many of these common questions tend to be variations on one of the following themes:

- 1) The common trend question: “Taking known seasonal changes and statistical power needs into account, is there a long term upward or downward trend in variable X?”
- 2) The common status questions: “Does variable Y exceed acute water quality standards instantaneously (one time)?” or “Does variable Y exceed chronic water quality standards often enough to be considered in violation of the state water quality standard?”

However, it is easiest to plan critical monitoring details if the general objectives in the central monitoring plan are rephrased into more detailed questions in each protocol narrative. A [QA](#) basic is that if the questions are sufficiently detailed, monitoring can more easily be planned in such a way that questions can be answered with the data collected. As monitoring protocols and SOPs are revised, it is important that the final more-detailed questions continue to make sense in comparison with summary discussions for [representativeness](#) and named [target populations](#) (or sampled populations) about which inferences will be made.

In the same general section where questions are being detailed, each protocol narrative should address the following checklist ([Checklist for Review of Vital Signs Monitoring Plans](#)) question: “Does the protocol narrative identify specific measurable objectives such as thresholds or trigger points for management actions?” When possible

the details of any such thresholds or trigger points should be fully explained in the protocol narrative.

An “ecological threshold” is generally said to be a rapid, non-linear change in system, while a “management threshold” is a point at which an action is necessary ([J. Gross, NPS, 2007 NARSEC Meeting Presentation](#)).

As was the case for questions, thresholds and trigger points should be discussed briefly in the protocol narrative and addressed in more detail in SOPs (such as the QA/QC or Data Analysis SOPs).

For example, are the thresholds of concern water quality standards? If the threshold or comparison benchmark is a water quality standard that already has some safety margin built in, managers may still want to know when a standard is being closely approached. Therefore, the magnitude of the change that needs to be detected in trend analysis detect should typically be smaller than the entire distance between current condition and the standard.

Is the threshold to be used a resource-collapse threshold value with no safety margin? If so, an even bigger safety margin would usually need to be factored into decisions about how big of a change needs to be detected. What units will the safety margin use?

Being able to detect a minimum detectable difference in original units (or an effect size expressed as a % of the magnitude of the standard deviation) of concern typically depends on variability of the data, sample size, alpha, and beta. These plus safety factors are input variables used to determine sample sizes and data [completeness](#) are usually covered in protocol narratives but should also be summarized in the QA/QC SOP.

The process for determining these issues is complex, so a much more detailed step-by-step by identifying quantitative desired conditions vs. current conditions, threshold levels, and safety margins can be found farther below in the [completeness](#) section.

Safety factors and thresholds should be covered at least briefly in the protocol narrative. If the network places details on these issues in the Data Analyses SOP or the QA/QC SOP, the network should also place a “point-to” marker in each protocol narrative so that readers can more easily find the more detailed discussions.

The protocol narrative should also summarize which questions and/or sites were selected to ensure monitoring of a 303d impaired water body or a very pristine water body that the park wants to keep that way. [WRD](#) has suggested that at roughly 2/3 of the sites should be in one of those two categories (see [Part A](#) of this guidance). What monitoring will be done to help answer GPRA reporting goals?

III. Document Vital Signs and Measures and How They Were Chosen

The protocol narrative should have a brief recap (or point to where the information may be found) on what will be measured and how vital signs and measures were selected. Was a set of neutral selection criteria used, such as those listed in Kurtz et al. (J. C. Kurtz, Jackson, L. E., and W. S. Fisher. 2001 [Strategies for evaluating indicators based on guidelines from the Environmental Protection Agency’s Office of Research and Development, Ecological Indicators](#) 1:49–60)?

The neutral criteria used in picking vital signs and measures should be summarized in the protocol narrative. Among the criteria listed (and discussed individually) in the longer version of [Part B](#) are:

- Select Parameters Useful in Answering Questions
- Select Parameters Relevant to Values to be Protected
- Select Parameters that are Logical Parts of Multiple Lines of Evidence
- Select Direct Measures of Specific Causes of Impairment
- Consider Parameters Commonly Measured By Other Groups (Ideally, Select Parameters Having Regional Data Sets Collected and Analyzed the Same Way-- Using Identical Protocols to Ensure Data Comparability).
- Consider Parameters Identified as Key Ecological Drivers
- Select Measures with Known and Moderate Variability at Reference Sites.
- Select Measures with Acceptable Minimum Detectable Differences (Within Acceptable Time Periods in Trend Analyses, Monitoring Design Sensitivity)
- Select Practical and Measurable Parameters
- Select Simple and Explainable Parameters
- Select Relevant Forms of Parameters
- Consider Composite Samples to Minimize Cost and Integrate Variability
- Consider Integrative Biological Response Variables [Especially Those Found to be Useful in Observed to Expected (O/E) Ratios]

A [QA](#) basic is that measures chosen should be helpful in answering stated monitoring questions. There should be a brief explanation in each protocol narrative of how the measures (typically level 3 Vital Signs) selected relate to both 1) values to be protected, and 2) desired conditions/ecological relevance. Which vital signs or measures were picked due to regulatory water quality impairment (303d lists, GPRA, etc. ([Part A](#))?)

Selected measures should ideally be simple and explained in plain language in the protocol narrative (see [Department of Water, Government of Western Australian 2004. Statewide Assessment of River Water Quality 2004 Methods](#) for a good example of plain language explanations for nutrients, standard water column parameters and DOC).

If comparison benchmarks are water quality standards or criteria, in what units are those benchmarks given by State or regional groups? A network may decide which specific subcategory and units of variables to measure based on the need to be fully comparable to the most relevant comparison benchmarks. Will the measures be total measures or dissolved measures?

Other things being equal, measures picked for long term monitoring should ideally not have:

1. Very poor measurement uncertainty (poor measurement [precision](#) and/or unacceptable measurement [bias](#)) or poor measurement sensitivity (usually [MDL](#)s or [AMS](#)), or
2. Extremely high true variability in the environment (from multiple different samples) at relatively un-impacted sites. This one directly effects the ability of a measure to be acceptable for overall monitoring design sensitivity.

The first of these relates to sensitivity, a QA/QC basic, on the scale of each single data point. The second relates to sensitivity at the scale of the entire monitoring design. If the minimum detectable difference is a 200% change in two hundred years, consider “changing something.” Among the options might be dropping the metric or measure or restricting the strata being monitored in time and space to strata with lower true variability.

For programs with limited monitoring budgets (= small sample sizes), excesses in either of these (or both) can prevent the detection of a true change of a magnitude of concern, or an optimal detection of a standards exceedance.

However, a counter consideration is that very important variables should not necessarily be thrown out just because they are extremely variable. Among the lessons learned by EPA in a major Mid-Atlantic exercise was that although certainly statistical criteria can be desirable, not all such properties guarantee a predictable association with a stressor of concern (such as human disturbance, L.S. Fore. 2003, [Developing Biological Indicators: Lessons Learned from Mid-Atlantic Streams](#), EPA/903/R-03/003).

A commonly stated goal of Principal Components Analyses is to see if just few components account for most of the variance in the data. If fewer variables can be used without much loss of information, it simplifies data analyses. For ecological indicators, a bigger issue than variance is typically responsiveness to individual stressors. Many State agencies have gone through efforts (EPA. 2007 [Biocriteria Homepage](#)) to determine the most important stressors, often using approaches suggested by EPA ([Stressor Identification Guidance](#)).

NPS can cite the lessons learned by the States and other Federal Agencies (for example, see the [Wadeable Streams Assessment](#), WSA) when documenting why measures or indicators were thrown out or kept. However, even when a goal has been to use fewer variables without much loss of information and also to use neutral criteria to identify measures particularly associated with certain degraded habitats, stressors, or impairment in general, the methods have not always been fully explained. For example, in the WSA, EPA concluded that “The most widespread stressors observed across the country and in each of the three major regions are nitrogen, phosphorus, riparian disturbance, and streambed sediments. Increases in nutrients (e.g., nitrogen and phosphorus) and streambed sediments have the highest impact on biological condition; the risk of having poor biological condition was two times greater for streams” (EPA, 2007, [Wadeable Streams Assessment](#) (WSA), EPA Publication No. EPA 841-B-06-002).

Those sound like a cause and effect types of conclusions, but not well explained in the WSA were:

- 1) The relative risk (RR) and other statistics used to get to those conclusions, or

- 2) The fact that neither correlation nor other strengths of association (such as RR) are the same as causation.

As is often better explained in human epidemiology (from which relative risk analysis was derived), the probability of causation (at any given single case) cannot be computed solely from the relative risk ([S. Greenland. 1999. Relation of Probability of Causation to Relative Risk.](#)). Concerning the relation of strength of association (such as a RR analysis) to causation or its role in causal inference, the two are related but far from the same (for more details, see Freedman 1999. [From Association to Causation](#), and Rothman, K. J. S. Greenland, and T.L. Lash, eds. 2008. *Modern Epidemiology*, 3rd. ed. Lippincott-Raven, Chapters 2 and 4). These concepts can be a bit counterintuitive at first, but even if we know a factor is a "general cause" (can and does produce the outcome in some cases) and even if we also know the amount by which it increases risk e.g., the relative risk), from that information alone we still cannot tell what the chance is that a given (specific-case) effect was caused by the factor (Sander Greenland, UCLA, Personal Communication to Roy Irwin, 2008).

Nevertheless, it was good that the WSA (op.cit.) looked at relative risk and also separately tried to use neutral criteria at multiple steps to help select important metrics. Although the WSA conclusion did not perform a true cause and effect type analysis of the kind detailed in the EPA [Stressor Identification Guidance](#), its relative risk conclusions were based on nationwide data and are still of interest.

The 2007 version of the WSA did not explain the statistics behind the conclusions, but Appendix A of the 2006 version of the [Wadeable Streams Assessment](#), explained at least a few aspects. For example, three pre-ANOVA screening criteria (low range, a noise to noise ratio they called signal to noise, and an ANOVA-specific F-test for means) were used to help reduce the number of invertebrate metrics. No ANOVA was then performed. Instead, the data with the remaining invertebrate metrics was analyzed with observed to expected ratio models, and then "relative risk" calculations were performed on selected (but not all potential) stressors. Although not explained in either the 2006 or 2007 versions of the WSA, in separate communications, the WSA authors have clarified that what was done for relative risk was consistent with the explanation in a later paper (Van Sickle, J., J. L. Stoddard, S.G. Paulsen, and A.R. Olsen. 2006. Using relative risk to compare the effects of aquatic stressors at a regional scale. *Environmental Management* 38, 1020-1030).

In considering such things, be sure to remember that unmeasured (or not considered) confounding or driving factors are also potential causes of less than optimal analyses. An important concept is that no amount of statistical sophistication can make up for missing the most important variables! As one potential example, in the case of the WSA, evidently low oxygen was not one of the stressors considered.

NPS protocol narratives should clearly and completely state why measures were chosen, including the neutral criteria used in the selection criteria. This should be more convincing than recounting that the turtle expert said to measure turtles or that a phytoplankton expert said it was a good idea to measure phytoplankton.

Designing a sampling event so that it provides optimal (direct or indirect) insight into the effects of potential drivers of change is an art. The best designs will estimate the

most useful parameters with the greatest validity and precision, and balance trade-offs wisely (J. Koopman. 1999. [The art of study design](#)).

Are parameters to be measured those also considered to be important indicators by other federal agencies or by the State? Are the parameters to be measured also deemed important by interagency groups giving lessons-learned or driver advice on the best things to measure in different habitats?

For example, for those networks monitoring lakes:

The Clean Lakes Program regulations (40 CFR part 35, subpart H) list the primary components that could be monitored to characterize the biological component of a lake system, including algal pigments, algal genera, cell densities, algal cell volumes, limiting nutrients, macrophyte coverage (by species), bacteriological components, and fish flesh analysis. The regulations do not specifically require monitoring for fish or macroinvertebrates (though mussels could be important in some areas), and also notably absent from the list are zooplankton (EPA 2006, Lake and Reservoir Bioassessment and Biocriteria Technical Guidance Document, Chapter 2: [Lake Biological Monitoring in USEPA, Local, State, Tribal, and Regional Protection and Management Programs](#)).

Wisconsin developed standard protocols for monitoring lakes to compare with State biocriteria in water quality standards. They did not use fish indicators either ([Executive Summaries of State Pilot Studies](#)). Nevertheless, some long term monitoring programs in WI do publish fish monitoring protocols, (really mostly some SOPs, rather than a more complete NPS-style protocol, see WI [North Temperate Lakes LTER Fish Sampling Protocol](#)).

Fish populations are prominently included by at least some groups monitoring lakes. For example, EMAP discussions include fish assemblage work in lakes (EPA. 1997. [Environmental Monitoring and Assessment Program Surface Waters Field Operations Manual for Lakes](#), section--1.3.2 Fish Assemblage discussion).

If measures are picked that are not used by other state or federal monitoring agencies, is there reason to believe that proposed measures will become more standard in the future? One example might be remote sensing of lakes, estuaries, and even [big rivers](#), for algal blooms, chlorophyll a, color changes, or various other estimates. Those on NPS intranet can see the NPS December, 2005 [aquatic remote sensing summary](#)).

Remote sensing will become more common in aquatic monitoring, and the NPS may like to consider partnering with other state or federal groups already doing remote sensing (for example, see [Minnesota Statewide LakeBrowser](#)) and [poster summarizing Remote sensing for VS monitoring for both terrestrial vegetation and SAV at Fire Island](#)). There are also a large number of remote sensing options for measuring temperature from airplanes, but some can also be hand held and used wading or in boats ([Thermal Imager Vendors](#)).

Always Measure Required Parameters

As explained in more detail in the NPS [freshwater](#) and [marine](#) core water quality white papers, several core parameters are required any time aquatic sampling is done:

For freshwater, required parameters include specific conductance (differs from conductivity by being temperature corrected), dissolved oxygen, pH, and water temperature. In addition, photographic documentation of the collection site (a minimum record of one digital site photo) is recommended. These are such basic vital signs that they are required even sample size is too low to establish trends. Some of these are almost required as normalization, correlation, or explanatory variables. For example, low dissolved oxygen is often associated with fish kills, and pH is required to interpret ammonia toxicity. High pH measurements may relate to low carbonic acid levels and high pH is often correlated with high chlorophyll and/or high nutrient levels and/or various photosynthesis issues. Algae tend to remove carbonic acid from the water as they photosynthesize, and the rate of removal can thus depend on the time of day as well as algae blooms

Although only [qualitative flow](#) is “required” for freshwater Vital Signs monitoring of streams and rivers, because flow is being discussed here, we will also discuss quantitative flow measurements in this same section. Since the concentration of so many water column parameters is so strongly influenced by flow, the [WRD](#) strongly encourages vital signs networks to measure flow quantitatively at the site being monitored. Flow is also a key to figuring out total load coming down streams, and is also sometimes needed to classify sites when analyzing site results using multimetric or observed to expected ([O/E](#)) models. It is best to measure flow using the most quantitative method that funding and logistics will allow. When practical, consider quantitative and relatively rigorous flow-meter methods used by agencies such as USGS (USGS. 2007. [Measurement of Stream Discharge by Wading](#)) and EPA EMAP (EPA 1998. [EMAP Wadeable Streams Manual](#), see section 6 on Stream Discharge). Whose flow data will be used for comparison? Choose SOPs that will produce comparable data.

Some additional NPS references on [flow measurement](#) and related issues such as trend analysis of flow data can be viewed on NPS computers having access to the NPS intranet. This is included on the NPS Share-point information sharing site Water Quality Monitoring Group Site, a site that also contains a variety of useful tools and a wealth of information on other stream monitoring topics. Other detailed guidance flow/discharge SOP and guidance documents, including [Techniques of Water Resources Investigations Documents](#) and USDA summaries are found in [Part C](#) of this guidance

For one example approach to developing QA/QC SOPs for flow, USGS Georgia has a 2005 QAPP for flow ([Open-File Report 2005-1246](#)), a document based on a 1995 generic template to be used to develop USGS State QAPPs. As expected, the Georgia QAPP is more complete than the earlier USGS generic template. The Georgia QAPP covers a bit on bias (“In order to minimize systematic errors, field trips are rotated to different personnel every 3 years.”). NPS monitoring networks could create a measurement quality objective for this type of bias (say, for example, a maximum percent difference of 10% or 20%) and the frequency could be more often. OFR Report

2005-1246 (op.cit) also contains some guidance on accuracy (sic, some of which seems to relate more to bias than accuracy), This same QAPP also contains good content on QA in general as well as practical recommendations for field methods. Further, it also covers some practical sense guidance related to the number of verticals used: “Measurement of discharge is essentially a sampling process, and the accuracy (sic, partly they mean representativeness and partly precision) of sampling results typically decreases markedly when the number of verticals is less than 25.” Although neither precision nor sensitivity are mentioned in the Georgia USGS QAPP, monitoring networks can easily develop common sense QC controls for precision (occasional repeat measures) and sensitivity (AMS could be calculated once in a while, even if only once every few years).

If flow is a main vital sign rather than a secondary measure (just done while getting primary measures), monitoring design sensitivity could also be estimated with minimum detectable differences (MDDs) after some years of quantitative flow data is collected. If no trends will be estimated for flow, this last step (estimating MDDs) is unnecessary. If flow is not the primary vital sign, but instead a potentially explanatory covariate, QC ‘completeness’ need not be controlled either.

For some stream applications (not for extremely low flows such as seeps), some in the NPS and USGS now use “The Flow Tracker ADV” made by [SonTek](#) (No Government Endorsement Implied). SonTek is a subsidiary of YSI (the ADV stands for Acoustic Doppler Velocimeter). There is a free self-training video that goes with the FlowTracker and an available application note discussing automated QC. See USGS 2004. [Policy on the use of the FlowTracker](#) and more recent USGS [OSW Hydroacoustics summary](#) for details and cautions. However, the latest electronic instruments are not the only option. There is still a place for the older mechanical measuring systems due to the fact they are less expensive to replace and are very reliable (if well maintained).

For seeps, an interesting recent summary is E. A. Adams, 2005. Determining Ephemeral Spring Flow Timing with. Laboratory and Field Techniques: Applications To. Grand Canyon, Arizona. MS Thesis, Northern Arizona University, available on the internet at <http://www4.nau.edu/geology/theses/adams2005.pdf>. This study used electrical resistance (ER) sensors to monitor spring-flow timing of South Rim springs. The sensor detects an increase in electrical resistance associated with drying events.

If quantitative meter-methods or stream gauge methods cannot be done to measure flow, the float method used by various State Agencies (for example see Arizona, 2005, [ADEQ Biocriteria Program Quality Assurance Program Plan](#) is adequate (and typically far better than having no quantitative flow results). The float method has also been recommended by some Federal Agencies (for example, see C. C. Harrelson, C. L. Rawlins and John P. Potyondy. 1994. [Stream Channel. Reference Sites](#) U.S. Forest Service Technical Report RM-245). The float method produces quantitative results which can be particularly effective when the data is accompanied by [QC](#) documentation for [precision](#) and operator-[bias](#) (to bound the magnitude of reproducibility differences between at least two observers). The float method is also used by many volunteer groups, and is said to be “actually superior in streams too shallow for meters (< 0.2 feet), or those with a flow rate below the level of detection of the meter” [EPA. 2003. [The Volunteer Monitor 15 \(2\)](#)].

When quantitative flow can’t be done at all, networks are encouraged to get flow data from nearby sites to indirectly gain insight. However, what is REQUIRED by [WRD](#)

in all cases is at minimum is at least a STORET-terminology-consistent qualitative assessment of flow “severity. The following is from D. Tucker. 2007. [Vital Signs Water Quality Data Management and Archiving](#)):

Choices	Description
DRY	No visible water in stream (typical of dry period for an ephemeral/intermittent stream).
NO FLOW	Discrete pools of water with no apparent connecting flow (at surface).
LOW	Base flow for a stream or flow within roughly 10% to 20% of base flow condition.
NORMAL	When stream flow is considered normal (greatest time that stream is characterized by this in terms of flow quantity, level, or general range of flow during a falling or rising hydroperiod, but above base flow).
ABOVE NORMAL	Bank full flow or approaching bank full (generally within upper 20% of bank full flow condition).
FLOOD	Flow extends outside normal bank full condition or spreads across floodplain.

Although not strictly required, it is also a good idea to include notes with water quality (and especially contaminants) datasets about when conditions reflected first flush (rising limb after a dry period) flows, since such conditions can influence concentrations of many pollutants.

Except for “low flow”, similar terminology could also be used **for lakes, ponds, reservoirs, or wetland water levels**, though the terminology is not now standardized in STORET. If enough networks agree on terminology, we could suggest new terminology for STORET. For example, the network might choose a rating such as the following expressed a % of bank full:

- Low - (<25% of bank full or perhaps lower quartile of conditions)
- Intermediate (or normal?) – [25% to 75% of bank full, or interquartile range (25/75%) of frequency of conditions?]
- High (or above normal?) - (greater than 75% bank full or in upper quarter of high conditions?)
- Flood Stage (overbank) - (greater than 100% bank full)

If a more complex lake, pond, or wetland rating system is used, something relatively simple (like the above) might still be used in addition to the more complex terminology. Remote sensing water level categories might be unique.

As another example, the EPA lakes habitat protocol and SOP guidance mentions lake levels several times and specifies the following SOP (EPA. 1997. [Environmental Monitoring And Assessment Program Surface Waters Field Operations Manual For Lakes](#), EPA/620/R-97/001):

- 1) Estimate the vertical and horizontal distances between the present lake level and the high water line.
- 2) The riparian habitat characterization includes riparian vegetation cover, shoreline substrate, bank type and evidence of lake level changes,
- 3) (Section 5.2.2.1.3): Bank type and evidence of lake level changes--Choose the bank angle description that best reflects the current shoreline that is dominant within your field of vision and 1 m into the riparian plot: V = Near vertical/undercut (>75 degrees, S = Steep; >30 to 75 degrees, hard to walk up bank; or G = Gradual, 0 to 30 degrees, easy to walk up). Estimate the vertical difference between the present level and the high water line; similarly, estimate the horizontal distance up the bank between current lake level and evidence of higher level.

If the waterbody is dry, other water column parameters like pH and conductivity cannot be taken, but recording the fact that the habitat is dry may be important to tracking changes in frequencies of flow or water level conditions. Changing stream flows, and the specific question “How many streams have had major changes in the size or timing of their lowest or highest flows since the 1930s and 1940s?” was singled out as an especially important national freshwater ecological indicator in the Heinz Report (The Heinz Center. 2005. [State of the Nation's Ecosystems](#)). [Phenology](#) aspects can also be important.

If the Cowardin et al (1979) wetland classifications for hydrologic regime (saturated, temporarily flooded, seasonally flooded, semi-permanently flooded, permanently flooded, etc.) are used to describe the type of wetland, that should be done in addition to rather instead of the instantaneous water level qualitative terms such as those above. In other words, it is still useful to know how full the wetland was when making water quality of aquatic biology measurements in a wetland.

For marine or estuarine monitoring, required parameters include ionic strength expressed as conductivity and as salinity, pH, dissolved oxygen, and water temperature. In addition, the following are required:

- Location standard coordinates [GPS on collection sites and also consult the Universal Transverse Mercator (UTM) grid; on USGS quad];
- Local time (indicating standard or daylight-saving time);
- Water depth and sample depth;
- Tidal stage (e.g. high, low, or mid-tide) and direction (ebb, flood or slack water),
- Estimated Wave Height.
- Flushing time

- Tidal range
- Habitat description

NOAA's National Estuarine Research Reserve System also specifies continuous monitoring for temperature, salinity/conductivity, pH, oxygen, and depth at four sites (same as NPS required parameters but NOAA adds turbidity) within each of its marine reserves. Three of the sites are in presumably relatively pristine sites and at least one site being a relatively impacted site. NOAA specifies continuous monitoring and also monitors turbidity at each site (NOAA. 2007. [Water Quality Indicators Measured by Reserves](#) Webpage).

Which Nutrient Measures to Monitor:

The short answer is measure whichever nutrient forms are covered by state water quality standards, but usually always include TN and TP. EPA notably recommends that states include the following for nutrient assessment: total phosphorus, total nitrogen, chlorophyll-*a*, and some measure of water clarity (e.g., Secchi depth or photometer for lakes and reservoirs and turbidity for rivers and streams, see EPA. 2001. [Development and Adoption of Nutrient Criteria into Water Quality Standards](#)). EPA suggested criteria (which would be useful for benchmark comparisons to VS monitoring results) for each of these parameters for both rivers/streams and for lakes (EPA 2007. [Ecoregional Criteria](#)).

NPS employees can access some misc. [nutrient tools](#) on the NRPC Sharepoint Site.

One thing to keep in mind is that the previous “limiting nutrient theory” notions, notably that:

P is always limiting in freshwater and N is always limiting in saltwater;

are greatly oversimplified. Different kinds of plant life can out-compete other plant life depending not only on ratios of N to P but also on availability of Sulfate, Carbon, Silica, Dissolved Organic Nitrogen, and many other factors. Also, water bodies are not perfectly well mixed due to stratification and other factors. For a good tutorial on why single-limiting nutrient notions are now considered outmoded and oversimplified see W. Dodds. 2007, [It's Not Just Phosphorus That Controls Trophic State in Fresh Waters](#), an EPA sponsored web-cast tutorial archived on a Tetra Tech interagency [N-Steps Website](#) (a site with many nutrient references and tools, most based on TN, TP, and Nitrate).

For nutrients like nitrate and other variables that vary predictably during a 24 hour cycle in response to changing sun energy, variability can often be brought down (making trends easier to detect) by sampling only in narrow index period of time during the day (just after dawn for example).

If there are no state water quality standards (for example in some estuarine habitats), and a network could only afford to monitor two nutrient parameters in surface waters, total nitrogen (TN) and/or total phosphorus (TP) will often be the best choices. Many states have TN and TP water quality standards, and these two are most often the most relevant forms for total loading issues (related to TMDLs and [national studies such as those done by USGS](#)).

An exception would be if there is much more existing regional comparable data for some other form (for example dissolved nitrate + dissolved nitrite rather than TN), then networks might want to seriously consider measuring these other forms in addition to (or perhaps even instead of) TN and TP. If one could afford to monitor four nutrients, then networks might consider monitoring total dissolved nitrogen (TDN), Particulate Nitrogen (PN), Total Dissolved Phosphorus (TDP), and Particulate Phosphorus (PP). Why? Because these four together give one more information and one can still get TP by adding TDP and PP; and one can still get TN by adding TDN and PN.

When one can afford to measure only two nutrients in surface water, measuring only dissolved inorganic nitrogen (DIN) and Dissolved Inorganic phosphorus (DIP, usually the same as Soluble Reactive Phosphorus – SRP) is often not as advisable as measuring TN and TP, since doing so gives one less useful information for either ecological relevance, total incoming load, or for Redfield Ratio comparisons (Walter Dodds, Kansas State University, Personal Communication, 2006).

However, there are usually exceptions to most rules of thumb and dissolved fractions (TDN, TDP, DIN, DIP, DON, and DOP) are more popular in coastal or marine monitoring. The Southeast Coast Network of the National Park Service has developed a rationale explaining why TDN and TDP are important indicators of their estuarine waters (Eva DiDonato, National Park Service, Personal Communication). Likewise, EPA's 2005 [National Coastal Condition Report II](#) (EMAP) rated coastal conditions as good, fair, or poor based on concentrations of DIN and DIP.

Some networks (Northeast Coastal and Barrier Network) have decided not to measure nutrients directly but instead to monitor eutrophication responses (chlorophyll, SAV, water clarity, etc.). This has some appeal since these measures integrate over time and that some (like SAV) may have less variability than water column parameters and has not been opposed by NPS [WRD](#), but keep in mind that recently some important documents [the [Wadeable Streams Assessment \(WSA\)](#) for example] have re-emphasized the importance of nutrients as ecological drivers even for things like benthic macroinvertebrates.

IV. Include Detailed SOPs for All Field and Lab Methods

A [QA](#) basic is that methods should be explained in detail. Exactly what will be done in the field and lab? Reproducibility and transparency are not only QA basics but also sound science basics. The amount of detail in the SOPs should be sufficient so that someone outside the NPS could duplicate the methods exactly.

As required by [Oakley et al. 2003](#), various NPS [WRD](#) guidance documents, and modern QA/QC conventions, all protocols should include method-detail SOP(s). For convenience, the method SOPs may be broken down into two groups: A field method SOP and a lab method SOP.

Together, the SOPs should fully explain the planned “response design,” the process and methods of obtaining a response at a site, once sites have been chosen (EPA 2006. EMAP webpage [Protocols for collecting data at sample sites](#)).

Consider Existing Protocols and Guidance from Other Monitoring Agencies:

The NPS Water Resources Division has consistently advised monitoring networks to consider using existing protocols and SOPs rather than inventing new ones. There is not need to re-invent wheels that have already been invented and work well.

In water quality work, the USGS and EPA and many states have existing protocols and SOPs. So a key question for reviewers of draft protocols is “Does the protocol narrative provide evidence that the monitoring network fully considered using existing protocols or SOPs used by state(s) or large regional monitoring programs? Some existing protocols and especially SOPs can often be adopted and be used as is. Using existing protocols and SOPs to the extent possible is a good idea not only because it saves time but also because it helps with regional data comparability.

Will USGS National Water-Quality Assessment Program (NAWQA) or state or EMAP protocols and SOPs be used? What detailed field and lab protocols will be used to get a response at the site?

Will parts or all of other federal and state guidance for monitoring and biocriteria in various types of habitats be incorporated into Protocol Narratives or SOPs? Prominent resources that have come to our attention so far include:

Wadeable Streams:

[EPA 1990. Biological Criteria: National Program Guidance for Surface Waters \(EPA-440/5-90-004\).](#)

[EPA 2000. Nutrient Criteria Technical Guidance Manual: Rivers and Streams](#)

[EPA 1996. Biological Criteria: Technical Guidance for Streams and Small Rivers, Revised Edition - EPA/822/B-96/001](#)

EPA. 1999. [Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates, and Fish](#) Second Edition, EPA 841-B-99-002

Design and Analysis specifics and examples for wadeable streams, defining target populations and rotating basin designs and examples ([EPA. 2007. Design and Analysis specifics and examples for Streams](#), EMAP).

The Maryland Biological Stream Survey (MBSS) includes at least some QA/QC aspects not only for chemical parameters, but also for many biological and habitat measures. Bias/accuracy accuracy goals are partly controlled by assessing the % of identifications of fish and herps that are correct (90% correct in the 2000 survey) and precision is controlled with a measurement quality objective of less than or equal to a [RPD](#) of 20% for invertebrate metrics like number of taxa or % tolerant ([Maryland, 2000. Maryland Biological Stream Survey Quality Assurance Report](#), Chesapeake Bay and Watershed Programs Monitoring and Non-Tidal assessment. CBWP-MANTA- EA-01-10, and a similar [QA report for 2001](#) which included aggregated IBI score duplicate RPDs). [Maryland biomonitoring protocols](#) are available for various types of organisms and habitats. Although the

Ohio protocols were early models in many ways, the current Ohio Volume III: Standardized [Biological Field Sampling and Laboratory Methods for Assessing Fish and Macroinvertebrate Communities](#) (available as a non-searchable PDF file) still is a bit short on [QC](#) control (instead of measuring QC bias, the Ohio fish Protocol simply expects 100% of the identifications to be correct).

An example of a detailed protocol for a response design for wadeable streams is the EMAP Wadeable Streams Manual, which includes many SOPs [Lazorchak, J.M., Klemm, D.J., and D.V. Peck (editors). 1998. [Environmental Monitoring and Assessment Program -Surface Waters: Field Operations and Methods for Measuring the Ecological Condition of Wadeable Streams](#). EPA/620/R-94/004F. U.S. Environmental Protection Agency, Washington, D.C.].

A diagram of typical EMAP placement of sites along a river (showing the alternation of left, middle and right collecting spots) is in the EPA 2006. [Wadeable Stream Response Design](#) homepage. Several NPS Vital Signs Monitoring Networks either have water or aquatic protocols and SOPs in the works based (at least partly) on the Western Pilot model or are considering using it as a model (including the Northern Colorado Plateau, Rocky Mountain, Northern Great Plains and Greater Yellowstone Networks).

Non-wadeable Streams (Includes Big Rivers and Great Rivers):

References with ideas for SOPs and protocols include:

[EPA 2000. Nutrient Criteria Technical Guidance Manual: Rivers and Streams](#)

For ideas and examples of using probabilistic surveys, see EMAP discussion on the EPA 2007 [Initial Monitoring & Design Approaches for Great Rivers](#) homepage.

For archived Great Rivers summary documents and newsletters, see EPA 2007 EMAP [Great River Study](#) Homepage.

Another broader resource is EPA. 2006. Environmental Monitoring and Assessment Program [Great River Ecosystems Field Operations Manual](#), EPA/620/R-06/002.

Flotemersch, J. E., J. B. Stribling, and M. J. Paul. 2006. Concepts and Approaches for the Bioassessment of Non-wadeable Streams and Rivers. EPA 600-R-06-127. US Environmental Protection Agency, Cincinnati, Ohio, http://www.epa.gov/EERD/rivers/non-wadeable_full_doc.pdf). This document explains many data comparability subjects for riverine monitoring covers large rivers, but the methods for comparability versus Measurement Quality Objectives (MQOs, including sensitivity, precision, bias, and precision) seem broadly applicable to smaller (wadeable) rivers as well.

Some agencies (EMAP and NAWQA) have utilized response designs that call for only a single site visit each year, usually sampled during a narrow index time period. This is typical for monitoring designs conducted over large regional areas. However, one can choose sites with a probabilistic design and then decide to sample the same sites (or with in the same general area or reach) more frequently. Details in a response design are driven by the objectives and questions to be answered by monitoring. Other misc. references relevant to developing large river protocols include the following (Sam Brenkman, Olympic National Park, NPS, Personal Communication, 2007):

Dunham, J., G. Chandler, B. Rieman, and D. Martin. 2005. Measuring stream temperature with digital data loggers: a user's guide. Gen. Tech. Rep. RMRS-GTR-150WW. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 15 pp.

Flotemersch, J. E., J. B. Stribling, and M. J. Paul. 2006. Concepts and Approaches for the Bioassessment of Non-wadeable Streams and Rivers. EPA 600-R-06-127. US Environmental Protection Agency, Cincinnati, Ohio.

Johnson, D.H., B.M. Shrier, J.S. O'Neal, J.A. Knutzen, X. Augerot, T.A. O'Neil, and T.N. Pearsons. 2007. Salmonid field protocols handbook: techniques for assessing the status and trends in salmon and trout populations. American Fisheries Society, Bethesda, Maryland.

Peck, D.V., J.M. Lazorchak, and D.VJ Klemm (editors). 2000. Unpublished draft. Environmental Monitoring and Assessment Program-Surface Waters: Western Pilot Study Field Operations Manual for Wadeable Streams. U.S. Environmental Protection Agency, Washington D.C.

Schmutz, S., M. Kaufmann, B. Vogel, M. Jungwirth, and S. Muhar. 2000. A multi-level concept for fish-based, river-type-specific assessment of ecological integrity. *Hydrobiologia* 422/423:279-289.

Estuaries/Marine and Near Coastal Areas:

EPA 1990. [Biological Criteria: National Program Guidance for Surface Waters. EPA-440/5-90-004](#)

EPA. 2000. [Estuaries and Coastal Marine Waters Bioassessment and Biocriteria Technical Guidance, EPA-822-B-00-024](#)

EPA. 2007 Webpage: [Design and Analysis specifics and examples for estuaries](#)).

NPS 2007. [Marine & Estuary Water Quality Discussion Board](#): those with access to the NPS intranet interested in marine and estuarine monitoring should consider joining this group developed and maintained by Eva DiDonato of the Southeast

Coast Network: To join, go to the intranet site and then fill in your contact information and start reviewing the site and posting new information.

[EPA 2001. Nutrient Criteria Technical Guidance Manual: Estuarine and Coastal Marine Waters.](#)

Wetlands:

Protocol, Methods, and SOP reference documents:

Except for lab chemical measures, most of these are notably short on QA/QC aspects:

1. [Biological Assessment of Wetlands Workgroup - Home Page](#)
2. [Wetland Bioassessment Fact Sheets \(1998\) - EPA 843-F-98-001](#)
3. For ideas and examples of using probabilistic surveys, see EMAP discussion [Initial Monitoring & Design Approaches for Wetlands](#). This document points out that defining the target population and availability of sampling frames, can both be somewhat problematic.
4. EPA. 2006. [Criteria Development Guidance Wetlands. Chapter 6 of the Nutrient Criteria Technical Guidance Manual](#) covers some QA/QC basics, but mostly for lab analyses rather than field measures.
5. EPA 2002. [Methods for Evaluating Wetlands Condition](#)
6. EPA 2002. [Methods for Evaluating Wetlands Condition #12 Using Amphibians in Bioassessments of Wetlands](#). This document notably recommends randomly selecting new surveying locations for each monitoring activity every year to avoid trapping biases and to take into consideration yearly changes in hydrology and plant communities. Precision is discussed mainly in terms of the magnitude of confidence intervals rather than in a [QC](#) sense. Bias is discussed mostly in terms of bias of collecting methods or gear rather than in a QC sense.

Lakes and Reservoirs:

Protocol, Methods, and SOP reference documents

- [Biological Criteria: National Program Guidance for Surface Waters \(1990\) - EPA-440/5-90-004](#)
- [Lake and Reservoir Bioassessment and Biocriteria: Technical Guidance Document \(1998\) - EPA 841-B-98-007](#)
- EMAP discussion of [Design and Analysis specifics and examples for Lakes](#)
- [EMAP SURFACE WATERS FIELD OPERATIONS MANUAL FOR LAKES](#) that includes many SOPs.
- EPA 2006. [Nutrient Criteria Technical Guidance Manual Lakes and Reservoirs.](#)

For lake or pond monitoring of amphibians, fish, and invertebrates, there are various helpful Website Resources outside of EPA:

For lakes WI publishes SOPs (See [North Temperate Lakes Long Term Ecological Research](#) website. For example for fish methods, first click on the "data" tab, then scroll down to "data protocols", then to "fish field sampling" to find the SOPs.

As another example, many states have standardized protocols and SOPs available, particularly for fish and benthic macroinvertebrates (BMIs). These State protocols are typically designed to assess compliance with biological criteria. State bioassessment and aquatic biocriteria contacts may be found by clicking on the applicable state on the map at the EPA [Bioassessment Programs](#) website. Among the states that already have detailed protocols and advanced narrative biocriteria are Idaho, Oregon, Arizona, Maryland, and Vermont. Most other states have some standardized protocols/methods and are at various stages of working on more advanced biocriteria to include in water quality standards.

One can search for existing protocols for amphibians and other groups of organisms using The National Biological Information Infrastructure (NBII) [Natural Resources Monitoring Partnership](#) Monitoring Protocols Library and Monitoring Locator System.

Consult NEMI on Lab Methods

In picking methods, a good first step is to scan methods and SOPs in the National Environmental Methods Index ([NEMI](#)) to get a relatively quick idea about which ones can achieve true method detection limit ([MDLs](#)) lower than all water quality standards or other comparison benchmarks or thresholds of concern. When possible, use methods that have acceptable MDL detection limits rather than RNGE—(range-defined) detection limits.

When possible, choose methods and labs where the semi-quantitative (MDL) detection limit that can be achieved is lower than the lowest water quality standard or other benchmark. Better when possible; choose methods and labs allowing the MDLs to be 1.6 to 2 times lower than comparison benchmarks. Best when possible, the MDL should be more than 3.18 times lower than the water quality standard or other comparison benchmark. The quantitative limit ([ML](#), minimum level of quantitation) detection limit is 3.18 times the MDL, so this is the ideal scenario where both the MDL and ML would be below all data comparison benchmarks, standards, or thresholds.

Getting the absolutely lowest detection limits (better and best examples above) is more important in some situations than in others. If monitoring networks anticipate that many of their measurements will involve very low level signals (low concentrations near the MDL detection limits), it is worth going to some trouble to find methods and labs that can achieve MDLs that are below the anticipated levels and comparison benchmarks. This is especially true for chemicals that can be a concern when present even in very low amounts and for nutrients in very pristine sites where nutrients are very low. However, for some analytes, methods or labs that can achieve detection limits that

low cannot be found. In other cases, all measurements are likely to be well up in the quantitative range. For example in farming or urban areas, nitrate levels in surface water quality are never likely to fall to levels near low-level detection limits. In this case, the lowest possible low-level detection limits may not be needed and the network could consider using methods and labs with higher detection limits if doing so reduced costs.

While one is quickly screening methods in [NEMI](#) to see if the method can achieve acceptable detection limits, it would be time-efficient to also check to see if the listed method [QC](#) performance for [precision](#) (usually as an relative percent difference --[RPD](#)) and [bias](#) (% difference or % recovery are sometimes listed for bias) are acceptable for project purposes. If acceptable detection limits, precision performance, and bias performance capabilities are not listed in NEMI or in the method itself, it is reasonable to consider other methods already having acceptable performance documented.

Put the Details in SOPs and their Appendices

Lab SOPs should detail exactly how everything is done in the lab. If a standard method from a state, USGS, or EPA is used, it should be written out or attached in its entirety and electronic copies should be archived in the database so that users can find out exactly what was done years from now. If no electronic versions exist, hardcopies should be archived and “point-to” notes in the database should give the location of storage. Method and SOP documentation should include measurement quality objectives ([MQOs](#)) for [measurement sensitivity](#), [Precision](#), [Systematic Error/bias](#) (bias is still wrongly called accuracy in some methods), and [blank](#) control bias. Many of the EPA methods are in [NEMI](#) and can be copied electronically.

If the agency (EPA, USGS, etc.) changes the method, will the NPS also change in the same way? Regardless of how this is decided, the detailed methods that the network plans to start with need to be detailed in the SOPs and archived for future comparisons. The SOPs should be detailed enough to allow third parties to reproduce the methods and to allow determinations of data [comparability](#).

Attach a QA/QC SOP to Each Aquatic Protocol: This topic will be mentioned again in much more detail below, in the subsections entitled: “[Why Document Quality Control?](#)” and “[Include a QA/QC Comparison Table for QC Topics](#)”. However, since [QC](#) methods are so often considered a part of methods, and since methods are being discussed here, the concept of attaching a QA/QC SOP to each protocol is first introduced here.

The QA/QC SOP should detail what will be done with data from samples that exceeded holding time requirements. Will such data be rejected or flagged? Data rejection and re-sampling so that newer replacement samples meet the holding times is usually the better option but flagging may be better than simply using the data as though it was high quality data. If flagging is chosen, it should be justified and flagging should be STORET-compatible as follows:

In cases where the sample exceeded the recommended holding time, and a decision was made to keep the data, the network can enter the data in STORET with a STORET remark code of "EHT", designating the condition that the "Sample or extract held beyond acceptable holding time." The data can also be

entered the same way in NPSTORET ([Vital Signs Water Quality Data Management and Archiving](#)). In NPSTORET, four fields over to the right is the Lab Remarks field where one can select "EHT" and/or other remarks/data qualifiers.

Examples of other details to be included in protocol narratives or SOPs rather than in central monitoring plans:

More details on sampling locations and method specifics.

For example, in chapter 4 of the central monitoring plan a table might say that chlorophyll *a* was the parameter to be monitored. SOPs with each protocol should document the rest of the details. For example, a field methods SOP might clarify that chlorophyll *a* is to be monitored using field water collection procedures of the [USGS field manual](#). A lab methods SOP might then further clarify that in the USGS national NWQL lab in Denver would do the work using USGS Schedule 1637 method. Alternatively, the lab SOP might specify another method (such as EPA method 445.0 or APHA method 10200H-4), was to be used. The entire method used should be copied and included as appendix to the appropriate SOP.

If flow or water level is to be recorded, will it be qualitative or quantitative?

What field instrumentation will be needed?

What pre and post season activities are required?

The field methods SOP should detail how will samples be collected and preserved, what containers will be used, and what maximum holding times were used. Unless otherwise justified, use holding time guidance in 40 CFR Part 136 to 136.3 and appendices.

Secondary Data Collection from Existing Data

It is not unusual for both the NPS and potential partnering agencies to have the generic problem of being short of funding for new sampling but also already having a substantial amount of data that has been collected but that has not been well summarized or analyzed for issues of concern to management or trends.

In some such cases, a high priority and proper use of NPS VS water quality funding may actually be to perform “secondary data collection” (one type of “re-sampling”) of raw data out of existing regional reports to find the data relevant to NPS needs. This is one type of “data collection,” but the distinction is that one is not going out in a waterway and collecting brand new raw data.

Such data can be periodically summarized with an emphasis on identifying hints of trends or threshold exceedances relevant to NPS management needs.

In typical such scenarios, the NPS would develop relatively short secondary data collection protocols and SOPs. These would summarize decision criteria for deeming

data comparable enough for the purpose of merging into statistics and/or using at all. For example, how different would [measurement quality objectives](#) for QC sensitivity, precision, and bias have to be before the data would not be considered comparable enough to merge into one statistical analysis? For general data usability, will only data with adequate QA/QC be considered useable, and how will adequate be defined? See definitions of useful and effective data in [Part B](#) (the longer version).

Other than that, the networks would also typically copy the protocols and SOPs used by other agencies to collect the data. In other words, there is still a need to archive all protocols and SOPs used by the other agencies, and put them in an appendix or other place that they can be found in future NPS Vital Sign Network sources (NPS. 2005. [Protocol Development Process NPS Vital Signs Monitoring Program](#)).

Those methods in the National Environmental Methods Index ([NEMI](#)) can usually be easily cut and pasted into NPS protocols, SOPs, or additional details in appendices. All data collected with Resource Challenge Money, even data “re-sampled” (secondary data collection) from other agencies rather than collected new by the NPS, needs to be in local network databases (to the degree that it documents data used in NPS analyses) as well as in National STORET, and data users need to be able to find the method details in the SOPs or associated appendices.

There have already been some success stories where small amounts of NPS funding combined with partnering from other agencies has resulted in the hope of producing products of great utility not only to the NPS but to the partnering agencies (see [Appalachian Trail Vital Signs](#) as one example).

In a similar vein, an example of a non-NPS effort to integrate data from multiple agencies was a State of Maryland effort that (although a bit short on data comparability documentation) has interesting GPRA or “report card” type summary graphics on multiple water quality and biological response integrator variables (Maryland. 2004. [State of Maryland Coastal Bays](#)).

V. Monitoring Design Summary in Protocol Narrative

The [QA](#) topics of overall monitoring, design, representativeness, and target populations are discussed together, because the way one assures representativeness is to name the target population and then design the monitoring to sample the target population in such a way that the samples obtained: 1) are representative of the target population, and 2) help answer previously identified questions.

By the time NPS VS networks are working on protocols and SOPs, many basics relevant to these topics should have already been summarized in chapter 4 of the central monitoring plan. The new task is to put additional detail about target populations and how representativeness will be assured in each protocol narrative.

Target and/or sampled populations can be summarized in a table in the protocol narrative. Additional detail on exactly what will be done to ensure representativeness could be placed in the representativeness section of the QA/QC SOP.

A domain is basically a synonym for a subpopulation, like a Sitka Spruce area grouping, which might be a domain but is not recommended stratum as it can change with forest fires, etc.

Representativeness

Whether representativeness is considered [QA](#) or [QC](#), a key question is: “How does the monitoring plan assure that the samples measured will be representative of the named target population”?

Representativeness is most often thought of a qualitative part of QA. One typically ensures representativeness statistically by having defensible monitoring designs, typically incorporating at least some randomness. This is suggested not only herein but also in generic (not just water quality) Vital Signs monitoring guidance documents (S. Fancy. 2000. [Guidance for the Design of Sampling Schemes for Inventory and Monitoring of Biological Resources in National Parks](#)).

Representativeness is a basic that needs to be discussed in a complete manner in every QA/QC SOP. It also needs to be discussed in less detail in the protocol narrative. Together the protocol narrative discussions and SOPs need to answer the following questions: “Representative of what?” These questions need to be answered even if USGS or other widely used protocols are utilized.

Department of Interior (DOI) and [Park Service information quality guidelines](#), as well as more generic NPS QA/QC guidance documents encourage “a high degree of transparency.”

One reason that defined target populations or sampled populations are compared to sampling designs and questions to be answered is to help insure transparency. In other words, don’t hint that your conclusions apply to all waters of the park when they really are only applicable to daytime only, late summer only, low-flow conditions only, riffles only, one stream only, or near one specific bridge only.

There is growing recognition that unless care is taken to ensure representativeness, data can be of little value, no matter how good the measurement performance is for [precision](#), [bias](#), detection limits, etc. In other words, ensuring data quality means not only insuring analytical quality but also sample representativeness of the target population given the questions to be answered.

A helpful and instructive example exercise on how hard it is to pick a representative sample based on “what seems right”, is found in the USGS exercise “[Can you select a representative sample?](#)”

Given what is known about variability in time and space, how will the sampling scheme insure that the values obtained will be representative of the [target population](#) being monitored ([Checklist for Review of Vital Signs Monitoring Plans](#))? If the answer is not in the protocol narrative, a statement should be made in the narrative as to where to find the answer. As one hypothetical example, the protocol narrative might state:

“Twenty five to fifty stratified random samples (or spatially balanced probabilistic-selected samples) will be collected from all flowing waters in the park. All flowing waters of the park, at all times of day and times of year and all locations, will have a chance to be selected, assuring representativeness to all flowing waters of the park. The target population and the sampled population are both “all flowing waters of the park” (see more detailed discussion in the representativeness section of the QA/QC SOP attached to this protocol).”

Although this would be a good example, most monitoring networks would probably reject something this broad due to cost and other practical considerations. A network might start with a very broad definition of the target population when first writing chapter 4 of the central monitoring plan. However, later when they start further developing protocols, SOPs, and monitoring plan [optimization](#) steps, they would probably then whittle it down to something more realistic. See more realistic examples for [copper and arsenic](#) below.

Target Populations versus Sampled Populations

The target population is simply the larger universe of all possible values (bounded in time and space) that one is sampling from and wishes to make statistical inferences (conclusions) about. This definition assumes the ideal situation where the target population and the population actually sampled are the same. Note to biologists: the “target population” usually does not necessarily mean a biological population in the sense that biologists often use the phrase, as a specific level of organization (contrasted to the higher “community” or lower “individual” scales of organization).

Many monitoring and statistical guidance documents state that a target population and a sampled should ideally coincide. For example, an OMB committee came to this conclusion and also stated that if there is a large set of units in the target population that has no chance of selection, the design is not a probability survey ([Federal Committee on Statistical Methodology](#), OMB. 1988. Statistical Policy Working Paper 15).

Most monitoring networks cannot afford to randomly sample all habitats at all times and in all places. Therefore, it is often useful to initially define target populations in very general terms (say all waters or all flowing waters or the park or network) and then later specify a more restricted “sampled population,” with inference only extending to the sampled population.

If monitoring networks decide to make a distinction between [target populations](#) and [sampled populations](#), (EPA, 2006, [Frequently Asked Questions - Survey Design](#) . EMAP), then it is important to define both in as much time and space detail as possible. The exact project-specific [sample frame](#) should also be defined, and also how often the frame will be re-done in long term monitoring (every 15 years, every 30 years?). It is really better to define target populations in terms of the larger population of potential values rather than in more vague terms such as the resource about which information is wanted.

For example, if no sampling is to be done in the winter or at night, the sampled population and sphere of inference or conclusions should not include night time or winter conditions.

In final reviews about whether or not what is proposed makes sense, networks need to compare the sampled and target population to the questions to be answered and extent of inference.

Copper Example:

For example, suppose the question is “does the concentration of copper in the water column in all flowing waters of the park ever exceed state chronic water quality

standards for aquatic life?” In this case, the network would typically need to consider when worst case conditions typically occur. If the network really wants to determine if there were ever any exceedances of the standard in any flowing waters anywhere in the park, then one would not want to restrict sampling to riffles only. Furthermore, if one wanted to know if copper standards were ever exceeded, then one would sample at night, the time where water column samples of copper tend to be the highest, not during the day. Fish and other aquatic life do not just live in the water during the day.

To answer this question, one must consider worst-case conditions. One would restrict sampling to short summer index period (for example, July and August low flow periods), if one already had reason to believe that is when copper concentrations would be the highest.

After considering the above example for copper, the network might realize that it was unrealistic considering budgets available and decide to make some new monitoring plan [optimization](#) changes in 1) target or sampled population, 2) extent of inference, and 3) the basic monitoring design. If the network really wanted to answer the question of copper standards exceedance, the discussion in the protocol narrative might include the flowing example.

1. The target population might still be “waters of the park,” but only if the network also defined a “sampled population” and only if the sphere of inference and conclusions are not to extend beyond that sampled population.
2. In the study design part of the protocol, the network might clarify that were stratifying by time of year, by flow conditions, and by night-only times for sampling. If the network or their advisors prefer not to call these restrictions strata, they can simply call them “response design details.”
3. Again, for emphasis, no matter what terminology is used, the key factor to be stated in the protocol narrative discussion of target populations, representativeness and monitoring design, is that sampling will, in fact, be done only during those restricted spheres of time and space, and that the extent of statistical inference (and the sampled population) will not extend beyond those restrictions.
4. The type of probabilistic sampling design should be listed in the protocol narrative (random, stratified random, or spatially balanced random hybrid designs such as the “[GRTS](#)” design. The type of design could be lined up with questions and basic approaches in a table in the protocol narrative. Some sort of probabilistic design would be needed if inference were to be made broadly beyond the areas and times immediately sampled.

Budgets often restrict the sampled habitats, sampled locations, and/or sampling times. The sampled population would seldom include all waters of the park (big rivers, small wadeable streams, lakes, wetlands, and ponds). Indeed, different protocols & SOPs would typically be needed for each of these radically different types of habitats.

Refine Monitoring Design and Representativeness Iteratively

The planning process is typically iterative with continual refinement, so when changes are made in the protocol narrative and SOPs, go back and make sure all of the related discussions in Chapter 4 of the central monitoring plan are still correct and consistent with the additional text.

For example, suppose a network initially named a target population as “flowing waters of the network.” Suppose further that sampling was in fact only going to be done in the daylight in low flow conditions during a July and August summer index-period, and only riffles were going to be sampled. The sampled population (and sphere of inference and conclusions) then includes only those potential values that could be measured during those very specific conditions.

As an example of how monitoring planning often proceeds in an iterative manner in the gradual fine tuning of monitoring details, consider representativeness. After thinking more about representativeness, the network might reconsider some of the “target population” details (discussed just above).

For example, would the network really be able to do the night-time sampling on a regular basis? Is that really the most important question to answer? The network might decide that it was just not realistic to sample at night. They might consider answering a question about daytime target populations instead, considering the limited monitoring funds available and other real-world practicalities.

Such changes are not unusual when monitoring networks are optimizing monitoring designs. Often monitoring networks have more vital signs, measures, and questions than they can answer with the budget at hand.

Again, [optimization](#) steps usually include throwing out vital signs, measures, questions or strata in time or space. Some potential monitoring approaches might be thrown out because they require night time sampling or other aspects considered impractical or dangerous. Others might be thrown out because some other agency is already covering the monitoring. After calculations of minimum detectable differences over longer time periods, still other measures might be thrown out due to excessive variability (even at pristine sites) and a resultant inability to find even large magnitude trends against the background of the high variability (high noise). Optimization might also include restricting the target population in time and space.

Consider Interagency Design Recommendations:

To help make sure that 1) objectives, 2) questions to be answered, and 3) monitoring design details; all line up with each other and with what will be done to assure [representativeness](#); we recommend that Interagency recommended tables (ACWI and NMQMC. 2006. [A National Water Quality Monitoring Network for U.S. Coastal Waters and their Tributaries](#)) be included in protocol narrative drafts. Such tables would be along the lines of the following (the following is just an example and would have to be modified according for Park and network-specific details):

Alignment of Objectives and Management Questions

Objective	Management Questions
1. Define status and trends of key water	What is the condition of the Nation's

quality parameters and conditions on a nationwide basis.	surface, ground, estuarine, coastal, and offshore waters? Where, how, and why are water quality conditions changing over time?
2. Provide data relevant to determining whether goals, standards, and resource management objectives are being met, thus contributing to sustainable and beneficial use of coastal and inland water resources.	Are strategies that protect or remediate water quality working effectively? Are we meeting water quality goals and standards?
3. Provide data to identify and rank existing and emerging problems to help target more intensive monitoring, preventive actions, or remediation.	What are the water quality problems? Where are the water quality problems? What is causing the problems?
4. Provide data to support and define coastal oceanographic and hydrologic research, including influences of freshwater inflows.	What research activities will help us to understand water resources and ensure they are sustainable?
5. Provide quality-assured data for use in the preparation of interpretive reports and educational materials.	All management questions require these data.

In the protocol narrative, networks are supposed to reiterate earlier questions and make them more detailed in time and space. The ACWI and NMQMC 2006 document (op.cit, above) also gives the following example of a Monitoring Network Design Summary. Something similar to this would be also be helpful in NPS protocol narratives.

Resource component	Purpose	Reporting unit	Number of sites per reporting unit	Total number of sites	Site Selection	Sample frequency	Sample interval
Estuaries	Condition of US estuaries	National & IOOS regions	50 per IOOS region	500 sites sampled per year	Probability-based design that will assure geographic coverage	Once per year	5 years (repeat year 1 sites in year 6)
	Condition of individual estuaries	Individual estuary	50 sites per estuary except for very small estuaries	1500 sampled per year (50 sites X 30 estuaries sampled per year)	Probability-based design that will assure geographic coverage	Monthly for physical and chemical conditions in water column; Once per year for biological characterization and sediment quality	5 years (repeat year 1 estuaries in year 6)

	Transport through estuaries	Individual estuary	15 sites per estuary	2235 (15 sites X 149 estuaries)	Distributed along salinity gradient from major river mouth to seaward outlet	Monthly for physical and chemical conditions in water column	On going
	Short-term variability	Individual estuary	2 per estuary	298 (these sites are subset of sites used for transport)	At two ends of salinity gradient	Continuous monitoring	Continuous

In the table above, the Southeast Coast Network NPS Vital Signs Monitoring Network and some other networks are basically using line two to estimate proportions impaired at individual parks (in different years) and the last line (short-term variability) to understand the [diel](#) component of variability at two sites per park representing extreme cases. Both of the tables above (modified as necessary, remove the lines or columns not applicable to the individual network) would be helpful for inclusion in protocol narrative text summaries.

Representativeness versus Diel Water Column Measures:

It is well known that oxygen, pH, and temperature (the core of our required parameters) tend to vary substantially in a 24 hour period in many shallow surface waters strongly influenced by sunlight energy. Less well known is the fact that concentrations of nitrates, metals, and many other water column parameters tend to do the same. In fact, it is more difficult to name example water quality parameters that never show diel signal changes in shallow waters than ones that do (see additional discussions in [Part B](#)).

Additional diel discussions, including lake discussions are available to NPS employees on the [NRPC Sharepoint diel site](#).

When dealing with measures that show strong diel signal strength changes, one needs to consider how sampling plan details need to be optimized to enable one to find long term trends. If sampling crews just happen to sample in the morning for a number of years and then a later crew just happens to sample late afternoons, a change in the data may simply reflect the changed time of sampling rather than a true environmental change.

Likewise, the target population is seldom all values that could be obtained during random sampling over 24 daily cycles, so the factor needs to be considered when naming target populations and trying to ensure the sample values will be representative of that target population.

As an instructive hypothetical example, let's suppose that not only is copper a potential issue at park, but so is arsenic. Suppose further that another equally high priority question was "Do water column concentrations of arsenic flowing into our linear (riverine habitat) park ever exceed water column samples for arsenic?" That would be an easier question to answer, because water column values of arsenic tend to be highest in the afternoon rather than in the middle of the night. Also, only one site would need to be monitored. The protocol narrative discussion might then be changed to reflect the following:

1. The protocol narrative would state that a “targeted” (“judgmental”) sampling design is appropriate to answer the question rather than probabilistic or random design. The question does not relate to all waters of the park and to answer it we only need to sample where the river flows into park jurisdiction
2. Likewise, the target population is no longer “flowing waters of the park.” Now it relates to potential daytime values that might be collected in only one location. The sampled population might be “water flowing into the park, where the river crosses into the park boundary (or close).”
3. In this hypothetical example, let’s assume that the network has no knowledge of one season being worse than another. The protocol might then state that 30 sampling dates (during the course of a year) will be picked randomly. If arsenic is worst case during a narrow seasonal index window of time, the protocol narrative should state that and specify monitoring will be done within that window of time.
4. To capture worst-case conditions for arsenic, sampling of the water column would need to be done in the afternoon only, when arsenic was likely to be highest in the water column.
5. The sampled population and extent of inference would therefore not extend beyond afternoon conditions. Also, since sampling will only be done in one location (where the river comes into the park), the extent of geographical statistical inference will not extend beyond that one location
6. In the study design part of the protocol narrative, the network might clarify that were stratifying by hydrograph limb period to try to take out most of the variability associated with many contaminants (especially metals, organics bound to soil or suspended particles, conductivity/Specific Conductance, or phosphorus compounds) that tend to spike during the rising limb of a storm event or snow melt event. Variability would tend to be lower during stable low flow periods, so one strategy to reduce variability (and therefore have a better chance of detecting a change of certain size) would to only sample (and to only infer about) the populations of values during stable low flow periods.
7. To further reduce variability, monitoring networks might also consider stratifying by time. For example it is common to sample only during narrow seasonal index periods say late summer only for example).
8. For trend detection for many parameters subject to large diel swings, another potential strategy is to stratify by time of day (hours after sunrise or before sunset) to try to take out most of the diel variability. The reason for trying to reduce variability is to enable detecting of trends of a magnitude of concern within a reasonable period of time. Most (shallow, sun driven) water-column parameters show diel variability in certain types of locations, especially pH, oxygen, temperature, chlorophyll, many dissolved metals, and nitrates. One sampling strategy that could be stated in protocol narratives is that sampling will be done on a diel basis at first and then later done in restricted periods of time to either get variability down or to capture worst-case time periods. Such details are typically part of response design SOP details. For metals and a few other “highly-pH-dependent” parameters, the variation is sometimes more dramatic in shallow flowing waters more heavily impacted by sun energy and diel pH changes, waters

- not well buffered, and/or sites having a relatively large proportion of the water column areas being sampled in contact with photosynthetic organisms (like algae, either benthic or phytoplankton) and sediments or suspended sediments containing metals. For diel patterns, see Irwin 2004. [Considering Variability When Looking for Trends](#) presentation and [Part B](#) for details. For other typical examples, see also USGS 2007. [Development and Research: Diurnal Metal Variations in Streams](#) gallery. Although it doesn't seem to cover diel issues, another general reference on risk from metals is EPA 2007. [Framework for Metals Risk Assessment](#)
9. There appears to be less information on diel cycles for water column nutrients and metals in large lakes and reservoirs than in streams. Here are few tidbits we have found to date:

Discharges from reservoirs can often be less variable than for streams, but such discharges (especially diel pattern discharges, but also just bottom water discharge) from large reservoirs can nevertheless change diel patterns downstream (see S. J. Hueftle and L.E. Stevens. 2001. Experimental Flood Effects On The Limnology Of Lake Powell Reservoir, Southwestern USA, Ecological Applications, [Volume 11, Issue 3 \(June 2001\)](#)).

Gaseous mercury diel patterns have been studied in reservoirs (Dill, C., Kuiken, T., Zhang, H., Ensor, M. Diurnal Variation of Dissolved Gaseous Mercury (DGM) Levels in a Southern Reservoir Lake (Tennessee, USA) in relation to solar radiation. The Science of the Total Environment, 357, 176-193, 2006). Several lake and wetland studies appear to indicate that highest water column mercury concentrations in shallow lakes and wetlands would be somewhere between noon and late afternoon, but it depends on local conditions as the peak could come earlier in some lakes. The loss of mercury from lakes is related to DGM concentrations in surface waters, wind speed across the air/water interface, and air and water temperatures (for more information see Siciliano SD, O'Driscoll NJ, Lean DRS (2002). Microbial reduction and oxidation of mercury in freshwater lakes. Environ. Sci. Technol. 36:3064–3068. [PubMed](#)) and various mercury publications (Krabbenhoft, D.P., J.P. Hurley, M.L. Olson, and L.B. Cleckner. 1998. Diel variability of mercury phase and species distributions in the Florida Everglades. Biogeochemistry 40:311-325; and Krabbenhoft, D.P., C.C. Gilmour, J.M. Benoit, C.L. Babiarz, A.W. Andren and J.P. Hurley. 1998. Methylmercury Dynamics in Littoral Sediments of a Temperate Seepage Lake. Canadian Journal of Fisheries and Aquatic Sciences. 55:835-844).

Diurnal cycling of oxygen, like pH cycling, is driven by sun energy, and variations are sometimes higher in eutrophic lakes. Slack water in shallow areas of lakes can increase algal and SAV activities, so it should not be assumed diel cycling is not in issue for both pH and metals in lake or

reservoir environments, it depends on many factors, and changes in pH do not universally guarantee big changes in dissolved metals. Other factors might include photo-reactivity and various biological processes controlling the movement of metals in and out of biofilms or other biotic classes.

Not only can dissolved metals concentrations change on a diel basis, but the species of the dissolved metal can also change. For example, Maest et al. 1992, looked at diel cycles of arsenic and iron REDOX species in Mono Lake, California, and found that during the day, dissolved iron was present as Fe^{2+} -- in the presence of dissolved oxygen and at pH 10 -- and as Fe^{3+} at night. The pH was not changing and the presence of reduced iron in the upper water column during the day could have been related to photoactive Fe-organic complexes. The authors did not see any obvious photo-reduction of arsenic, but dissolved arsenic did change from arsenate (AsV) to arsenite ($AsIII$) below the chemocline in the lake (A.S. Maest, S.P. Pasilis, L.G. Miller, and D.K. Nordstrom, 1992. REDOX Geochemistry of Arsenic and Iron in Mono Lake, California, USA, In Water-Rock Interaction VII, Y.K. Kharaka and A.S. Maest (Eds) Balkema, Rotterdam, pp. 507-511).

Representativeness versus Tidal Cycle Signals:

Strong tidal cycles can make the task sorting [diel](#) signals out from tidal signals very difficult. The most ideal scenario would be able to sample at the same time of day during a specific time of the tidal cycle, which is very difficult or impossible to pull off. It depends of the questions one is trying to answer, but sometimes the tidal signal can overwhelm the diel signal, so the Southeast Coast NPS VS network has settled on methods recommended by NOAA's National Estuarine Research Reserve program guidance. The NOAA SOP specifies taking nutrient and Chl samples between 3 hours before low tide & low tide, to at least consistently sample the same water body (the estuarine waters without the marine influence). In one sense this stratifies time to cover limited and consistent tidal stages, and so also limits the target population but probably increases the chance of detecting trends (Eva DiDonato, NPS, Personal Communication, 2006).

A contrasting approach, which would increase the universe of inference but would probably also increase variability (complicating trend analyses), would be to randomize more completely in both time and space. One advantage of a probabilistic survey is that [as long as one has covered both space and time adequately with sufficient sample sizes to ensure representativeness of the true underlying population being sampled, one will presumably eventually be able to demonstrate that the full range of conditions have been covered. In this manner, the extremes from both tidal and [diel](#) factors will be covered in more random probabilistic designs, but at the expense of increasing variability.

Representativeness in Wadeable Streams:

Back to the somewhat easier freshwater wadeable stream scenarios, if the network in our hypothetical example decided that they want to keep an eye on copper trends, even though they can't practically sample at night, they might decide to look at trends rather than water column exceedances. They might further decide to sample copper in sediments rather than the water column. If they understood local variability enough, response design details might call for other sampling restrictions. For example, in one small stream in Yellowstone, it was discovered that variability could be reduced by sampling metals in sediments only in low flow late summer conditions and only in the sediments of low gradient riffles, where variability is lower than in the water column or in other sediment microhabitats (such as backwater pools).

Trying to get the variability down is done so that the monitoring network can have a reasonable chance to detect a change of concern (say a 30% change in means over a stated time period (say 1 year or some alternative time period) without collecting hundreds of samples every time they went out. See Yellowstone example in [Part B](#). Some networks (and many states) use hybrid sampling plans that include both 1) targeted sites (such control sites or historically sampled bridge sites) to answer site-specific or other limited inference questions and 2) probability-selected sites that allow for broader inferences to larger areas of the park or waterbody. Such hybrid designs are often good compromises but sometimes tend to stretch funding even further and make getting 25-50 samples per year in each park more difficult. Why do we need 25-30 samples? See EPA explanation ("[Why a sample size of 50](#)").

At least one NPS network (Southeast Coast) has tentatively proposed to solve the problem of getting enough samples for a good estimate of a proportion by taking 25-50 probabilistic samples per year **at one park only**, and rotating to other parks in future years. To better understand temporal variability at each park each year, the network has also proposed to use continuous monitoring of NPS required parameters at two sites in each park: 1) A representative impacted site and 2) A relatively un-impacted (or relatively pristine) site. Other factors are also sometimes used (for example salinity in estuarine sites) to bound two extremes.

Contrasting relatively pristine with relatively impacted has the advantage of book-ending intervals that would define extremes at the two kinds of sites. This approach for continuous monitoring of a few key parameters at key sites was inspired by a consistent approach NOAA uses in its Marine reserve monitoring program. NOAA does continuous monitoring for our NPS required four parameters (temperature, salinity/conductivity, pH, and oxygen) and also measures turbidity at four sites within each reserve. At least one site monitored is a relatively impacted site and the others are relatively pristine (NOAA, 2007, [Water Quality Indicators Measured by Reserves](#) Webpage).

The idea of combining probabilistic sampling (to answer large area questions) with fixed site continuous monitoring at two extremes (to answer questions about short term variability and trends at fixed sites) has now been recommended by interagency groups. In fact it was highlighted in the 2006 recommendations of ACWI and NMQMC. In this case, salinity was used to define the extremes rather than impacted and less impacted (ACWI and NMQMC, 2006, [A National Water Quality Monitoring Network for U.S. Coastal Waters and their Tributaries](#)).

If All Sites Were Selected With a Judgmental Approach

If absolutely no randomness is to be involved in picking sites to sample, the rationale should be justified on a network-specific basis. Why not? With what logic would a targeted design assure [representativeness](#), what is target population, and why do the pros of the targeted design outweigh the cons? Typical pros given include limited funding, the need to continue long term historical trend data at specific sites, and just the complexity of balancing changing target populations and changing (optimal) sample frames.

Note: The phrase “sample frame” refers to the list or map that identifies every unit within the target population of interest, a physical representation of the target population. Such a list is needed so that every individual member of the population can be identified unambiguously. As explained in the EPA definition of the sample frame, the individual members of the target population whose characteristics are to be measured are the [sampling units](#).”

In human sampling surveys, a phone book is often the sampling frame. The phone book contains all the names in the target population, assuming the target population is all the names in the phone book. Assuming that the phone book is supposed to be representative of all the humans in a town brings in a potential bias problem, since not all humans have land-line phones and of the humans that do have such phones, not all choose to be listed, and of those listed, some groups of people are home or answer such calls more than other groups of humans. An even more severe problem in water quality sampling is that if one considers all potential water quality concentrations that could be measured in that general area and that general time-period as “the target population” one never really has a complete list of all of those values. If one had all the values, one would have a census and sampling a smaller set of values to infer to the larger population would be unnecessary. But we seldom do have this, and what we tend to have instead is a general location in space of (say for example, all perennial streams in a certain area) and a general location in time (daytime, say July and August). A further complication is that over a long period of time in long term monitoring (say 100 years), both the target populations and the optimal sample frames (optimal to sample those changing target populations), will be changing too. No one would consider using a phone book that is 100 years old as a sample frame. These types of real world issues would tend to necessitate eventually (periodically) re-randomizing from new sample frames, or at least periodically including some percentage of new sampling locations from new sample frames.

If a network chooses to make all sites judgmental or targeted sites, with absolutely no randomness at all, they still need to address [representativeness](#) and [target populations](#). In the absence of convincing evidence to the contrary, the target and sampled populations, as well as the extent of inference, will all be limited to those sites sampled only.

Even if the stream near a bridge is selected in for long term monitoring, there are things that can be done to approve the quality and usefulness of the data. These might include:

1. Sampling far enough way from bridges to minimize bridge-effects (deicer salts, dust, vehicle pollutants, trash, changed hydrology, etc.) to help with representativeness with this general area of the stream and not just the (often unusual) conditions right at the bridge.
2. Once the network has picked the area to sample (say upstream of a bridge, possibly in riffles only), they can still pick exactly where to sample in the riffle randomly, using simple stop watch field randomization (see [Part B](#)).

Alternatively a network may decide to use guidelines such as those used by the USGS National Water-Quality Assessment Program (NAWQA) [sediment quality collection protocol](#) to maximize data comparability with NAWQA. NAWQA specifies collecting sediment samples in low flow periods only (to reduce seasonal and flow driven variability), and they specify compositing samples from different microhabitats (within and among different zones) to get an average for a reach and to reduce variability driven by habitat type (and to make the samples more representative of a larger area).

For water column parameters, the USGS Field Manual gives decision rules about how well mixed a river has to be to allow for a single grab sample at midpoint rather than compositing several samples from a cross section:

"If profile values of pH, conductivity, temperature, and DO differ by less than 5 percent and show that the stream is well mixed both across the section and from top to bottom, a single measurement point at the centroid of flow can be used to represent field-measurement values of the cross section" ([USGS Field Manual, Wilde and Radtke chapter 6](#): Section 6.0.2.A. page INFO-11,). Essentially, this is a bit like having a measurement quality objective of differing by no more than 5% for [precision+](#) (NPS term for nearby but different replicate samples).

Using such guidelines is fine and often has the advantage of helping achieve data comparability with other agencies with other large regional data sets. Again, one still has to address the question: "representative of what?" In other words, the USGS method only assures that a single sample is representative of a cross section at that single location in the river at that point in time. It does nothing to insure that the single sample is representative of locations up or downstream or at other times.

Whatever understanding one has of terms like strata, the target population, the sampled population, and zone of statistical inference, all such phrases should be clearly defined and be transparent to readers.

If the agency believes that sampling a cross section of surface water will be representative of some areas upstream that have not been sampled, how far upstream, and based on what data? At minimum, comparisons should be done before starting monitoring to check all such assumptions, and these should be repeated occasionally over the years. Otherwise, such beliefs will be based strictly on speculation rather than on any actual data. It would take many such comparisons (more than most networks can afford)

to convince most statisticians and survey design experts. Thus the most common practical alternative is simply to state that inference will not extend beyond the particular site location (and/or times) that had a chance to be monitored.

Causation

Documenting causation (not a requirement in NPS VS monitoring) is difficult to prove without active manipulation. Inside labs, only one variable is typically changed at a time, and the rest are kept constant. This makes it much easier to tease out cause and effect stressors. Outside in the environment, countless variables (temperature, rain, wind, clouds, solar storm-induced changes in solar radiation, etc.) are changing all the time, often in unknown or less than fully-understood ways. So to get at potential causation, one usually approaches like those in EPA's stressor identification document, which summarizes strength of evidence analyses using multiple lines of evidence (EPA 2000. [Stressor Identification Guidance](#)).

Recently EPA has developed the more user-friendly [Causal Analysis/Diagnosis Decision Information System \(CADDIS\)](#). This system provides a pragmatic guide for determining the causes of detrimental changes and undesirable biological conditions observed in aquatic systems. A Caddis summary table on [typical types of evidence](#) helps explain the concepts.

In a sideways reference to how tough it is to prove causation in complex outdoor systems, a newspaper columnist (Sullivan, J. 1995. Field and street. Chicago Tribune, August 1995, quoted in Stow et al. 1998, [Long-term environmental monitoring: some perspectives from lakes, Ecological Applications: Vol. 8, No. 2](#), pp. 269–276) stated:

“Right now the Great Lakes are like a very poorly designed experiment set up by an incompetent scientist who figured that 600 or 700 variables would be just about right for his protocol.”

However, Stow et al. 1998 (op cit.) clarify that “Despite the size and complexity of the Lake Michigan ecosystem, the many confounding factors, and relatively noisy data, a sufficiently large sample size ($n = 589$) provides a basis for choosing among alternative models that represent different mechanisms and have different management implications.”

None of the above prevents monitoring networks from thoughtfully placing sites in such a manner that hints relative to causation can be obtained, but remember that stressors tend to change over time and this is long term monitoring.

Stratification

If a monitoring network has decided to stratify, they should avoid using strata where variability characteristics (in time and space) are not well understood or are likely to change appreciably during the monitoring period. Keep in mind that this is long term monitoring, so eventual change is more likely than for short term projects. It is often safer to stratify by factors that change less frequently or dramatically (often geological or physical factors), or to handle timing and detailed space issues in the response design

rather than in a more general monitoring design in chapter 4 of the central monitoring plan.

Typical patterns of variability and typical patterns of response to various single stressors say nothing about other patterns of variability and other responses not considered. In other words, a stratification pitfall is that it is often harder to group sites into homogeneous groups, especially for multiple stressors, than one first thinks (S. Urquhart. 2000. [Adapting a Physical Habitat Protocol](#)).

Often monitoring planners need to think through collection details carefully in order to get variability down to magnitudes that would allow detecting a reasonably small change without hundreds of samples. For streams, if variability characteristics are understood well enough, one can stratify by habitat types (such as low gradient riffles only, or runs only, or snags only). Monitoring groups are sometimes able to document that the variability reducing aspects of stratification outweigh the disadvantages. The stratum description might be qualified to take into account the changing environment of streams, climate change, etc. One can also specify index collection time periods in either stratification decisions or in response design decisions.

A useful document on the typical need for stratified random sampling of outdoor environments (classified as non-experimental studies of uncontrolled events) was provided by Schwarz 1998 (Chapter 3 in [Statistical Methods for Adaptive Management Studies](#)).

Again, if a monitoring network chooses to handle such details under the response design documentation rather than calling it stratification or handing them in stratification steps (as part of the spatial monitoring design), one can put the needed details in individual protocol narratives and SOPs.

GRTS and Similar Approaches for Assessing Status

GRTS is the acronym used for generalized random-tessellation stratified monitoring designs. In some ways, GRTS is a hybrid between random and systematic designs. Each site has at least some probability of being selected. In other words, the probability is never zero. GRTS designs are designed to ensure a degree of spatial balance that can often be lacking in purely random sampling.

Disclaimer: There are other spatially-balanced survey designs that may be as good as or better than GRTS, and even when considering GRTS alone, there are competing software programs that claim to be implementations of GRTS. Furthermore, there appears to be at least some controversy about which programs are a narrowly-defined GRTS approach and which are different enough they should not be considered GRTS but rather some other way to achieve spatial balance in a probabilistic survey design. No government endorsement of one particular approach is implied here. The only reason we are focusing a bit more on GRTS in this section is that several NPS monitoring networks have proposed its use in long term Vital Signs monitoring

Spatially-balanced samples are random samples. They just don't happen to be simple random samples or systematic grid samples. Just as with simple random samples,

a combination of unequal weighting of probability of selection (plus, in some cases, stratification to reduce variability) can result in more intensively sampling certain targeted sub-regions and help negate the likelihood of getting sites where personal safety or access are a big problem.

NPS generic Vital Signs monitoring guidance points out that while the following should be avoided

Judgment sampling, using "representative" sites selected by experts,

two other attributes are helpful when deciding how to sample the named target population of interest "Probability samples occur when each unit in the target population has a known, non-zero probability of being included in the sample, and always include a random component (such as a systematic sample with a random start)" (NPS. 2006. [Sampling Design Considerations: Where and When to Sample](#)). Probability sampling and random components help assure [representativeness](#).

For example, the NPS. (2007) document entitled [Rocky Mountain Network Monitoring Plan](#) states that: "The spatially balanced samples produced by GRTS are more representative than those produced by other probability designs"...and "When site replacement rules are strictly followed, the representativeness of the final sample is still guaranteed."

Both GRTS and simple random sampling involve probabilistic strategies. No matter how such issues are decided, all decisions should follow a careful and documented (in the protocol narratives) thought-process. GRTS can be used with or without stratification.

GRTS probabilistic designs can be good choices when done right, when sample sizes per year are high enough, and when all aspects are logically defensible. A big draw to such designs is the ability to infer to larger areas and not just to those being sampled, while still largely avoiding (though unequal weighting) unsafe sites, sites too logistically difficult to sample, etc. As long as sample sizes per year or other logical sampling unit are high enough, GRTS designs can produce status-friendly and GPRA-friendly proportions (% of stream miles impaired, % of flowing water achieving water quality standards, % of flowing waters where an index results in a rating of excellent, etc.).

A few additional summary remarks on GRTS are made here, but most monitoring networks considering using GRTS or similar variants would be well served by having an applied statistician familiar with spatially balanced applied survey designs help sort out the details. Many potential pitfalls are too complex to easily and completely summarize herein.

One thing to keep in mind is that GRTS and other probabilistic designs help one decide **where** to collect samples **in space**, whereas panels and other revisit schemes help one decide **when** to sample **across time**. Changes in magnitude as well as variability can be driven by both changes in time and space. In long term monitoring, if both places and timing are changing, then one has to pay attention to how such changes in both time and space might complicate or influence long term trend analyses, and whether or not the (target) populations being sampled are changing or staying the same.

For those with advanced quantitative skills, more information on [Technical Information for Implementing Designs](#) (including unequal weighting, spatially balanced

designs, the four step process of implementing GRTS in general, and [download software](#) for S-plus and R, are available on EPA EMAP websites).

Although only [SPSURVEY](#) software does GRTS samples for points, lines, and polygons, two other programs do spatially-balanced sampling using somewhat different approaches:

S-DRAW, available from WEST Inc., is a software implementation of GRTS for point (finite) spatial populations. According to the West Inc. website, S-Draw is meant to serve the purpose of “[GRTS for the Average Joe.](#)” The Heartland NPS VS monitoring network (HTLN) used S-Draw in its [invertebrate monitoring protocol](#) (available on NPS computers on the intranet only). Specifically, “to draw the GRTS samples, main stem sites were weighted by stretch length” and HTLN also “employed the reverse hierarchical ordering option, which assures that any contiguous set of stretches will be spatially balanced.”

ArcGIS RRQRR uses an alternative spatial-balance algorithm. It also doesn't utilize the same hierarchical selection process typically used for GRTS to provide spatial balance. Due to the differences, it should perhaps be thought of as a related but different way to get spatial balance. Nevertheless, the CSU website claims that: “The Reversed Randomized Quadrant-Recursive Raster (RRQRR) algorithm is an implementation of the Generalized Random Tessellation Stratified (GRTS) algorithm” ([Spatially-balanced sampling using RRQRR](#) based on [Theobald, D.M. and J.B. Norman. 2006. Spatially-balanced sampling using The Reversed Randomized Quadrant-Recursive Raster algorithm: A User's Guide for the RRQRR ArcGIS v9.1 tool](#)).

A potentially helpful resource (for those with quantitative backgrounds) is additional information from related presentations that were presented at the San Diego National Vital Signs Meeting in 2006 (See presentations by [Schweiger and Urquhart](#)).

Will a Probabilistic Monitoring Design be used for Status or Trends?

The original GRTS-selected (or simple random selection process) sites help one to be able to infer to sites not sampled for **short term status**, but less clearly for **long term trends**, especially if revisit and new-site addition details are not thought out carefully in light of statistics that will eventually be used to analyze for trends. If one next repeatedly goes back for the next 100 years to only those exact spots first selected for sampling during the first year, then one's ability to infer more broadly will (eventually) be harder to defend. That first GRTS draw might just so happened to have picked sites not optimally representative (i.e. biased) of the then-existing target population, a issue that re-randomizing or doing additional GRTS draws over 100 years would at least partly help correct. The first GRTS draw may become even more or less representative over a long period of time.

If one were trying to accomplish both goals (status and trends), how would optimal designs look in typical NPS Vitals Signs monitoring scenarios? There is no single right

and wrong answer, but whatever the answer is it needs to be logically justified and explained in each protocol narrative.

However, for one example of how it might look, consider the following scenario. A network might decide to visit one park per year [Southeast Coast Network (SECN) example], or one habitat type per year (say similar high altitude lakes at more than one park, [Klamath Network (KLMN) example], with a sample size of say 40, so that would could make valid conclusions (such as percent considered impaired) about status in the one-year time period. In the SECN example, the next year, they might sample another park. In the KLMN example they might visit another habitat other than high altitude lakes.

In either case, let's say the network attempts to get a sample size of 40 each year. Then when revisits are done say 3 years after the first sample collections, instead of re-randomizing and getting 40 new sites or simply revisiting the exact same 40 sites again, a network might decide to revisit 25 to 30 of the exact same sites (for trend analyses) along with 10-15 newly randomly selected survey sites (so that one is at least slowly getting new randomized sites relevant to broader inferences to a possibly changing target population). Then every 15 years (or other logically developed time frame), new GRTS draws would be made from a revised sampling frame (reflecting changed conditions), so that each year after that the 10-15 new survey sites would be randomly selected from the latest sampling frame. Why 10-15 new sites per year? There is no single right answer (others could be logically justified), these are just hypothetical examples. The National Agricultural Service (NASS) replaces 20% of their sites every year, just as one comparison.

Introduction to Probabilistic Designs for Long Term Trends

One important issue is how long it will take for an original randomization or GRTS draw, and the sample frame site selection was based on, to "wear out." This is not a simple issue. A changing target population (and changing optimal sample frames) might result in the need to re-randomize and/or change the sample frame.

GRTS and similar probabilistic survey designs that also involve rotating panels do not have a long track record for monitoring long-term trends in natural resources, and statistical analyses for long term trends can become pretty complex due to some of the factors discussed in this section, reminding all of the need to consult an knowledgeable applied statistician).

Not to be ignored is the fact that over long time periods, target populations tend to change. This is one reason that some GRTS experts tend to favor a split panel type design that rotate new sites in at several different time scales. Since target populations are logically changing in very long term monitoring (and since an optimal sampling frame would also be changing) some consideration should be given to possibly re-randomizing (or doing another GRTS draw) periodically, perhaps every 15 years or so, though the optimal time period might be shorter for some vital signs than for others (Tony Olsen, EPA, Personal Communication, 2007).

Ideally one would have some criteria to trigger the need to redo things, such as "re-stratify when 30% of the sampling sites are in a new vegetation type", or "resample

when the known set of sites (e.g., wetlands) changes by 25% " or something along those lines.

Over 100 years, not only would the target population change, but also the optimal sampling frame. Some areas may become more impacted than now, while other areas may become less impacted. Some streams may dry up, while others may change to higher flows. Some streams that are now reservoirs may be restored to flowing free. Some perennial streams may become dry. Some channelized streams may be restored to more natural patterns. As logging and other roads are removed from wilderness, some easy access sites may become more difficult. Some smaller streams that now show up on maps may become shopping centers. So eventually it will be necessary to do another GRTS draw or re-randomize to get representative samples of changing target populations.

The [Southwest Alaska Network Monitoring Plan](#) (Phase III) [Chapter 4](#) addresses some related "common sense" tests in rotating panels, although not in a water monitoring context: "An important consideration when choosing a revisit design is its ability to retain a representative sample across time. A sample that is initially representative may lose this quality if there are changes or shifts in population numbers or other attributes during later time periods that are no longer captured by the original sampled units. These shifts across time could be induced by natural changes (e.g., habitat succession), anthropogenic actions, or a combination of both. If large shifts are not expected to occur or if the membership design is spatially balanced enough to adequately capture any shifts, the best revisit design to detect trend is to repeatedly sample the same plots across time, all else being equal. However, repeated visits to the same units could potentially have a negative impact on the response, such as trampling in vegetation monitoring plots, which would introduce bias".

For short-term status monitoring, GRTS allows inference to a broader target population not sampled. If sites from that first GRTS draw are repeatedly revisited year after year, they become fixed site designs over time. In this scenario, sample size estimations are based on paired sampling, which generally demands smaller sample sizes for the same amount of statistical power compared to sampling where a new random sample is selected each year. Thus, inference to the broader target populations based only on these fixed sites, can be harder to justify over long periods of time, since in subsequent years when one is always going back to the same sites, the site selection basically changes to a fixed-site scheme. Over extended periods of time, it also becomes less likely that those originally selected sites remain representative of a changing target population, some parts of which had no chance for selection during a randomization done years earlier.

If the primary goal is to determine one-time status rather than trend, selecting a new set of sites for each time period is best. However, if the primary goal is to determine trends and estimate the trend magnitude, it is much more efficient to revisit sites. Change between two time periods at any given site based on a revisit becomes equivalent to "pairing." At least some sites need to be revisited to determine trends, but also to get hints about whether or not the target population may be changing so drastically that it has essentially become a different target population. Adding to the overall complexity, in long term monitoring, realism demands consideration of changes in an optimal sample frame, and how such changes impact all conclusions. So how do we design monitoring so that we can make conclusions about both status and trends? To be able to say anything

about the current status, one must have sufficient sample sizes (usually a minimum of 25, to reduce the confidence intervals about proportions, and often also about means, to acceptably small magnitudes. To accomplish both goals, it is generally best to pool resources within a given field season and sample the population(s) of interest adequately and forgo re-sampling the population in consecutive field seasons. The re-sampling should include both re-visits (these essentially become index or fixed sites) as well as some new randomly selected sites (survey sites). Bringing in newly selected survey sites on a regular basis is helpful at assessing larger target populations, but can also complicate statistical analyses needed to determine long term trends (Andrew Merton and Scott Urquhart, Department of Statistics, CSU, Personal Communication, 2007).

A common question in monitoring surveys is the use of temporary or permanent monitoring sites. For example, should permanent water quality sampling sites that are re-measured over time, or temporary sampling sites that are re-randomized at each time be used? Many of the concerns are similar to those for repeated sampling designs discussed earlier. Permanent plots give better estimates of change over time because the extra plot-to-plot variability caused by bringing in new plots each year is removed. However, the costs of establishing permanent plots are higher than for temporary sites, and the first randomization may lead to a selection of plots that have some strange characteristics. Of course, if the measurement process alters the sampling unit, new plots will have to be selected for each survey. A compromise solution is a rotating panel survey, where only a part of the sample is changed at each time point. In large, complex, long-term designs with multiple objectives, permanent plots are often the preferred solution since no survey design is optimal for all objectives and the objectives change over time (C. J. Schwarz. *Studies of Uncontrolled Events*, Chapter 3 *In Sit*, V. and B. Taylor (editors) 1998 [Statistical Methods for Adaptive Management Studies](#), B.C. Min. For., Res. Br., Victoria, BC, Land Manage. Handbook No. 42.).

The issue of how to calculate variance relative to sample sizes needed for trends analyses can be complex and [discussed separately below](#).

Why go to all the trouble to sort out the complex issues? Because this is long term monitoring, and we want to get it right for both status and for trends.

Helpful references for those wanting to delve into GRTS and other spatially-balanced variants, and complex timing issues pertinent to long term monitoring, include the following references:

Stevens, D. L., Jr., and A. R. Olsen. 1999. Spatially restricted surveys over time for aquatic resources. *Journal of Agricultural, Biological, and Environmental Statistics* 4:415-428.

Skalski, J. R. 1990. A design for long-term status and trends monitoring. *Journal of Environmental Management* 30:139-144.

Kish, L. 1965. *Survey Sampling*. John Wiley & Sons, New York.

Kish, L. 1986. Timing of surveys for public policy. *Australian Journal of Statistics* 28:1-12.

Kish, L. 1987. *Statistical Design for Research*. John Wiley & Sons, New York.

Kish, L. 1988. Multipurpose sample designs. *Survey Methodology* 14:19-32.

Binder, D. A. 1998. Longitudinal surveys: Why are these surveys different from all other surveys? *Survey Methodology* 24:101-108.

Holt, D., and C. J. Skinner. 1989. Components of change in repeated surveys. *International Statistical Review* 57:1-18.

Kasprzyk, D., D. Duncan, G. Kalton, and M. P. Singh. 1989. *Panel Surveys*. John Wiley & Sons, New York.

Theobald D.M., Stevens D.L. White, D., Urquhart N.S., Olsen A.R., Norman J.B., 2007. [Using GIS to generate spatially balanced random survey designs for natural resource applications](#), *Environmental Management* 40 (1): 134-146

GPRA and Proportions

Proportions are potentially useful for GPRA and other management and reporting goals. One caution: Beware of (or at least look closer when encountering) small sample sizes when estimating proportions. Keep in mind that sample sizes should be 25-50 to estimate a proportion well and that any proportion estimated for sample sizes below 25 is a big red-flag (see EMAP explanation "[Why a sample size of 50](#)" and additional discussion on [proportion size calculations](#) below).

A typical problem for NPS VS networks is that they often cannot afford (the optimally defensible for a proportion estimation) 50 aquatic samples per year in a [GRTS](#) design unless they use other generic VS funding to supplement water quality funding. Like many states, many networks do not want to put all their funding into a GRTS design but instead they often favor a hybrid design. Networks often desire to monitor at least a few targeted sites for long term continuity or to answer site-specific or resource-specific questions. Most networks also want to measure more than one aquatic variable and/or different variables at different types of sites.

Will the Information be Useful to Management?

A key question is what kinds of data would be of most interest to management? One reason for going to the trouble to think through the issues presented herein, including estimating minimum detectable differences and the need for at least some QC, is to avoid the following:

Too often past monitoring has resulting in filling up file cabinets and/or databases with data that (even if one tried) could not readily be used for resource management or trend analyses purposes.

Would a superintendent and Park resource managers be more interested in results from a sampled-population stretched over five years, or would that superintendent be more interested in how things conditions contrast between in wet years vs. dry years, cold years vs. hot years, or high flow times vs. low flow times. ? To protect the resource, superintendents may need to manage the resource differently in some types of years vs. others.

Another related issue relates to internal data comparability. Are the data sets from longer time frames (five years for example), comparable enough to be combined into one sample? Over five years, bio-technicians and other personal (and equipment) are likely to change, resulting in measurement bias changes. One might then have two different kinds of results for the same sample (higher results when one staff member did the measuring, and lower results when a different staff member measured the same sample).

Over longer periods of time, variation magnitude may also change for similar reasons. Both variation and summary statistics like averages would also be apt to change due to changes in the target population being sampled. These types of changes would be more apt to happen over a five year period than in one season, and would make it harder to defend that we logically have only “one sample” and not more.

As one example, let's say that one can afford only 30 samples per year. Would it be better to take 10 samples at each of 3 parks during each of 3 consecutive years, or would it be better to take 30 samples at each park each year and rotate through the parks over each 3 year period. Likewise, would it be better to take 5 samples at each of 6 parks during each of 6 years, so that one could eventually get 30 samples per park, or would it be better to take 30 samples at each park each year and rotate through the parks over each 6 year period?

Answer: if the answer one is seeking is a proportion (% stream miles impaired, etc.) sample size needs to be 25-30 to result in a credible calculated proportion. Accordingly, it would generally be better to get 30 valid samples from one logical sampling unit (one park, or say small backcountry lakes at two similar parks) in one year. That way, one at least has a believable (the confidence interval about the proportion is small enough to be credible) proportion for any given year. Then 100 years later, if one is comparing proportions over the 100 year period (say comparing proportions from wet years to proportions from dry years), at least each proportion being used is a credible stand-alone entity. In other words, it would be better to have credible proportions for logically relatively homogenous strata or Parks each year. If proportions are to be estimated based on composite data from multiple years, try to keep the years down to no more than two or three, since anymore than that tends to stretch credibility more and more. In other words, if a two year period is chosen, there is a better chance that those two years would be similar (dry years say), then if the one was compositing information from 6 years (Andrew Merton, CSU, Statistics Department, Personal Communication, 2007).

States that require relatively few samples (for example, 1 per month, or 4 per year for metals, both for two years) for the purpose of gauging compliance with water quality standards may be exceptions. In that case, there are enough samples for regulatory compliance simply because the state says that is enough.

Some federal programs also have regulatory-defined statistics that are required to be used by definition, regulation, or provided guidance. For example, in RCRA and

CERCLA assessments of contaminated soils (which would be analogous to sediments in the aquatic environments), [upper confidence limits \(UCLs\)](#) corresponding to 80-95% confidence are sometimes used not only for precautionary principle estimates of means (either parametric or nonparametric) but also for standard deviations or variances.

Regardless of state or federal regulatory “minimum” requirements, NPS resource managers may recognize the need for more samples thoughtfully placed in space and time to be more fully representative of the full range of conditions in the environment. For example, given that many metals vary diurnally, seasonally, and spatially, are 4 metal samples per year in a stream or reach logically enough samples to ensure scientific credibility (for example, to represent the full range of conditions in a representative way)? Usually not, and resource managers are usually interested in the true conditions and not just regulatory status. Only when the true condition is known can resource management be done in an optimal way.

Does It Still Make Sense?

A potential complication encountered by other networks using [GRTS](#) or other probabilistic designs, has been that lumping values from different years together to eventually get a big enough sample size may not always be logically defensible. On one hand, lumping five years of data might help cover a fuller range of conditions better than single years.

On the other hand, very small sample sizes (always problematic or at least worrisome in statistics) can be a fatal flaw, especially when combined with inattention to timing and spatial issues. If one only takes 30 samples from a very large area (such as a whole park or whole network) over one five year period, could one stand up in court (or even in front of a superintendent) and say with a straight face that the full range of conditions had likely been captured with our 30 samples? Although the fact that samples had been sampled randomly is perhaps even more important than sample size, having a reasonably large sample (that is representative of the full range of conditions) is also crucially important.

In the above example of 30 samples, the resulting proportion based on five years of lumped data may not be fully or optimally representative of the target population one was trying to protect. What would the target population or sampled population be? Whatever is chosen should pass last minute reality-checks of logic, defend-ability, explain-ability (keep it simple is optimal there), and common sense.

In the highly variable universe of water quality, it may be hard to logically defend the notion that five annual samples, of sample size six each (from a very large area) is in fact “one sample” and the right (or optimal) sample to estimate a proportion or average that would be truly representative of the named target population. Due to changed conditions, it might be easier to defend that there are in fact, five valid samples (not one). In worst case scenarios, [GRTS](#) plus excessively small sample sizes and inattention to timing and important variability-reducing response design details (such as sampling low gradient riffles only or snags only, or only in short “index” time periods) may produce data that is so variable and so anecdotal (small sample sizes) that it may not be useful for many (if any) purposes. This is one key reason why so much water quality data collected in the past has not been useful for management purposes.

A related potential complication is that [composite sample power analyses](#) must be handled differently than normal. This is not a fatal flaw by itself, but must be dealt with in defensible ways.

An important reason past data has too often not been useful relates to trend detectability. Five year averages estimated from extremely variable data might make it difficult to detect even big changes from one five year period to the next, or to detect longer term trends. Again, our monitoring design should produce data that is useful for resource management decisions and useful to answer stated questions.

Timeliness is another issue to consider. Resource managers may not consider conclusions that they get only once every five years to be timely enough. Superintendents have sometimes wanted to detect a change of less than 50% over one year. For certain rare or important biological resources, superintendents have sometimes not wanted to lose 50% without knowing it after one year, let alone after five years. This may be even more worrisome if the estimate is questionable due to small sample sizes.

Reporting data and QA/QC summaries (but not conclusions on trends or water quality exceedances) annually is necessary and helps. However, if after 5 years, meaningful summary statistics (means, medians, proportions, water quality standard exceedances, % meeting acceptable condition index scores, etc.) cannot be calculated, that would typically be a problem. Likewise, if after 10-20 years, if even true and substantial trends could not be detected because of study design flaws (often including inadequate sample sizes), resource managers and other data users will probably not be well served. Species which are legally protected or even locally rare would be difficult to manage based on conclusions once every 5 years. Again, they might disappear between conclusions. In such special cases, there may be missed opportunities for management, and resource managers may have very little warning about declining populations. With shorter intervals of monitoring, and credible sample sizes, there is a better probability of detecting trends or bad conditions in time to develop and implement management strategies to avert losses.

After Revisions, Go Back and Optimize Related Sections

The adaptive management and iterative way that monitoring planning should proceed is so important, that it is mentioned again. If Chapter 4 of the central monitoring plan discussed several different options for different monitoring designs and discussed factors that nothing in the protocol is designed to address, these different tracks need not be given substantial discussions in the protocol narrative. Instead of lengthy discussions of discarded or minor tracks in the protocol narrative, just provide a table with summarizes the option chosen (for example a probabilistic design to answer questions requiring random designs, or a sentinel/historical site design to answer questions that infer to trends at that one site only). If there are ancillary research questions of interest the network, they could be mentioned (related to potential future funding) but they need not be given much space in the protocol narrative. It is important to clearly separate sidelight or discarded strategies or questions rather than leaving the impression that the monitoring design, protocol, and SOPs chosen can answer more questions than it can. In other words, the protocol narratives should be focused and all related section sections (Chapter 4 and the SOPs) should be consistent with each other and with the protocol

narrative as iterative revisions are made. Readers will note that iterative changes are a consistent theme in this guidance, not only in basic monitoring designs, protocol narratives, and [representativeness](#), but also in the steps covered next:

QUALITY CONTROL (QC):

We have already explained that [QA](#) relates to a system of steps (including some qualitative ones) that are done to ensure quality in an overall systematic planning and project management process. For contrast, QC includes quantitative (measurable) performance characteristics for data quality indicators like measurement [precision](#), measurement [bias](#), measurement [sensitivity](#), and (for chemical measures only) [blank](#) control bias.

Accordingly, typical definitions of QC usually emphasize measurable Performance-Based Measurement Systems (PBMS). Such PBMS QC basics usually include performance and data acceptance criteria (EPA. 2006. [Guidance for the data quality objectives process. EPA QA/G-4, EPA/240/B-06/001](#)).

The next few sections also cover Completeness and Comparability. In documents of other agencies, these two are sometimes covered in [QA](#) sections, sometimes in QC sections, depending on the authoring agency and specific document. But all seem to agree that [sensitivity](#), [precision](#), and [bias](#) are QC topics virtually always accompanied by quantitative measurement quality objectives—[MQOs](#). EPA has used the related-phrase data quality indicators (DQIs) in the following manner:

EPA's 2002 QA/QC document (G-5) includes DQI descriptions for include [precision](#), [bias](#), accuracy, [representativeness](#), comparability, [completeness](#), and [sensitivity](#). The document also makes it clear that the word accuracy should only be used for controlling precision and bias in combination, usually based on reference materials and/or spikes (with larger samples sizes than two, since two would give one only one-way bias and not a good long term estimate of accuracy, Table D-3, Appendix D, In EPA. 2002, Guidance for Quality Assurance Project Plans ([QA/G-5, EPA/240/R-02/009 December 2002](#)).

EPA has clarified that “Quantitative DQIs” include precision, bias, and sensitivity, whereas “Qualitative DQIs” include representativeness, comparability, and completeness (EPA 2000. [Introduction to Data Quality Indicators](#)). EPA has also clarified that that completeness is a combination of quantitative and qualitative control (EPA. 1998. [EPA Guidance For Quality Assurance Project Plans](#)).

Why Document Quality Control?

The need for a separate QA/QC SOP was first introduced above in Chapter 5, since QC documentation details are often considered part of methods. However, at this point in the progression of topics that need to be covered in protocols and SOPs, we are turning out attention away from general qualitative [QA](#) topics and towards more specific and quantitative QC topics, so more detail on each QC topic is provided below.

In general, there is increasing consensus that point estimates are not optimal for most purposes, and in fact are no longer considered acceptable for many purposes. Just as it is now considered to be good form (and more scientifically defensible) to express uncertainty on average values with a confidence interval rather than just reporting a mean value as a point estimate, so it is now considered necessary to express uncertainty about each single data point.

The NPS [WRD](#) agrees with the [NIST](#) and [ISO](#) national and worldwide scientific consensus (as explained for the US by NIST, see N. Taylor and C. E. Kuyatt. 1994. [Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results NIST Publication TN 1297](#), that:

1. No (single) measurement is perfect. Each is an approximation, and
2. Individual measurement data points are not complete unless accompanied by a statement about the uncertainty of that approximation.

QC samples allow one to estimate the amount of uncertainty about each data point. As will be explained in more detail, uncontrolled measurement processes, which typically are not accompanied by any QC checks, are no longer acceptable. Just as confidence intervals express the uncertainty about a mean of many data points, an AMS interval can express uncertainty around each single data point.

One difference between Quality Assurance – [QA](#), and Quality Control—QC, is QA tends to be controlled qualitatively (translates in Vital Signs Monitoring to protocol narratives and in other parts of the overall monitoring plan), whereas QC tends to be documented and controlled **quantitatively** in QC SOPs. QC helps put quantitative boundaries on how imperfect the measurement process (relevant to each single data point) is allowed to be.

Thus, if a reading on a pH meter is 7.0, is it really best thought of as 7.0 ± 0.2 (which is probably fine or at least within project objectives) or is our confidence in the reading so low that it should really be honestly thought of 7.0 ± 3.0 (this much uncertainty would typically not be acceptable). And if the reading at one site is 7.0 and the reading at another site is 7.2, do we believe that the change is really reflective of the environment being measured in the two places really being different, or is just a reflection that our pH meter is not measuring very accurately, so that 7.0 is not really different than 7.2? Likewise, if one biological technician estimates “percent embeddedness of cobbles” at a site as 20%, and another technician estimates the % at the exact same spot as 40%, we immediately suspect that measurement observer [bias](#) is not being well controlled.

Without such controls, we have no way to estimate how badly the measurement process is performing. Measurement uncertainty could be extremely large (or suddenly change) and we wouldn’t even know it.

In modern science, QC performance results help us document that the measurement process has been kept “in control” within reasonably small limits. Before explaining this further, we will first answer the following question:

Why do we need to quantitatively control the measurement process for measurement [sensitivity](#), measurement [precision](#), and measurement [bias](#)?

The short answer is that most states and regulatory agencies require us to do so as part of a required quality assurance project plan (QAPP). Some states also have credible data laws that address these issues. Many regulatory processes and even many modern databases will not accept data without QC and other metadata documentation.

Even if no outside entity is “making us” do so, there are logical reasons why we should control and document the performance of QC data quality indicators:

Scientific Credibility: The scientific community has known since the 1930’s that measurement processes need to be quantitatively controlled for both [precision](#) and [bias](#) to produce credible data (Newman, M.C. 1995. Quantitative Methods in Aquatic Ecotoxicology, Lewis Publishers, Boca Raton, FL., p. 282). Reproducibility is not only a QC basic, it is a “sound science” basic. Unless one documents measurement performance characteristics, it will be very difficult for another party to reproduce the result independently. So documenting QC controls and results is simply part of documenting sound science in today’s world. Although one need not go overboard with QC, one needs **“To do something!”** to maintain QC (doing nothing to control QC is no longer a viable option).

Legal credibility: For similar reasons, without QC performance documentation, it would be difficult to defend our data if attacked on the basis of not only scientific credibility, but also legal credibility

Multipurpose Needs: It is expensive to collect data, so to the extent possible, data collected should be credible for multiple purposes and to multiple agencies (many of whom require QC documentation). This tends to be particularly relevant to regulatory goals (which usually require QC) and associated GPRA goals.

Long Term Usefulness: QC performance results help insure that the data is useful for a longer time period, including use in estimating trends. We are planning long-term monitoring, and controlling and documenting QC indicators gives our data a better chance to be considered credible in future years. In future years, we would not want our data thrown out because someone had then decided that all data without full quality control documentation was unacceptable and would not be used. This is already happening more and more often, and the tendency to do so can be expected to become more common. If our data cannot be used for its intended long term purposes, our monitoring funding will have been wasted, and/or budget cuts in monitoring would be easier to justify.

Method Change or True Trend Change? In long term monitoring, method, SOP, and staff changes are inevitable. Documenting changes in measurement performance after such changes is therefore even more important than for short term projects. Documenting QC performance of the old versus the newer methods helps one determine whether or not a change in values was the result of a true change in the environment or whether it was the result of a change in the measurement process. Thus, good QC documentation of measurement changes

that [bias](#) scores upwards or downwards makes finding long term trends in a defensible way possible. For more details, see Section XII ([Include a Cumulative Measurement Bias SOP](#)) herein).

Data Interpretation: We need QC results to be able to interpret data in an optimal and common-sense manner. For example, we need low-level measurement [sensitivity](#) results to understand whether or not the analyte is present and how big of a measured change is believable as a real change (rather than a random error in the measurement process). Poor [precision](#) is normal when getting close to the low-detection-limits of measurement sensitivity (usually within 2-3 times the [MDL](#) or AMS sensitivity limits). In field biology estimates if bird or amphibian calls are weak, the “signal” is not strong, the noise tends to be getting close to being as strong, and normal signal to noise ratio theory becomes applicable. However, if we don’t know what the detection limits or other measurement sensitivity limits are, it is much harder to interpret the meaning of the precision QC results.

Adaptive Improvements in Monitoring: Programs that have instituted improved quality control over the years see their QC scores improving, even for field measures (USGS 1998. [Summary of the U.S. Geological Survey National Field Quality Assurance Program from 1979 through 1997](#)). The corollary advantage to agencies using these kinds of checks is that if QC scores suddenly go downhill, the agency would know to make corrections until the situation was corrected. If there were no checks, an agency would not even know the measurement process had deteriorated.

Old style Peer Review is not enough: For those who might still believe that journal-style post-project peer review is all that is needed, remember why that solution (although certainly one helpful step) is not complete by itself. Among the reasons: 1) it is too late to change the design or outcome, 2) scientific journals that have studied their own peer review processes have found glaring inadequacies, 3) many projects that have been published in peer reviewed journal articles (unlike many government grey literature documents) have no documented QA/QC at all. A 2006 introduction to these problems is in the background introduction to the [First International Symposium on Knowledge Communication and Peer Reviewing](#). A similar [2007 summary](#) of some of these issues continues with these same themes, including why typical post-project does not work well and is typically based on false assumptions. Among the references prominently cited in both these summaries were:

Chubin, D., and E. Hackett. 1990. [Peerless Science: Peer Review and U.S. Science Policy](#). Albany, N.Y.: State University of New York Press.

D. Kaplin. 1995. How to Fix Peer Review, 1995. *The Scientist*, Vol. 19, Issue 1, Jun. 6.

J. Ziman. 1982. Bias, incompetence, or bad management? *The Behavioral and Brain Sciences*, 5 (2), pp. 245-246.

Internal Requirements: QC documentation is required by NPS [WRD](#), the [generic VS checklist](#), needed to fully comply with spirit of the DOI information quality guidelines, and needed for input fields in modern data bases such as STORET or NPSTORET. NPS (WRD) requires that all water quality data collected by the Vital Signs aquatic monitoring be put into STORET, and STORET metadata fields call for [QC](#) performance information).

Include a QA/QC SOP and Comparison Table for [QC](#) Topics

Each protocol narrative should include a QA/QC SOP that documents what will be done to control and estimate the magnitude of:

Measurement [Sensitivity](#) (Usually as [MDLs](#) or [AMS](#)).

Measurement [Precision](#) (Usually as [RPDs](#))

Measurement [Systematic Error/Bias](#) (Usually as PDs)

Measurement [Blank](#) Control Bias (If present above MDL magnitudes, this is another contributor to total measurement bias). Blank control is usually handled with separate blank [QC](#) samples in chemical lab analyses, see details below).

[Completeness](#) Goals: Although this relates to a [QA](#) topic at the monitoring design level rather than a strict [QC](#) goal, since the goal is quantitative, it may also be helpful to include the completeness objectives as part of a QA/QC summary table. This helps put project quantitative goals all in one place.

Generally QC Measurement Quality Objectives and Frequency of QC samples for each of the above should be no less stringent than requirements of the State. If data will be compared to other large Federal Data Sets (USGS NAWQA, EPA CERCLA or EMAP, FWS, NOAA, etc.) then blank control requirements should also be no less stringent than the other Federal Program whose data will be used for comparison with NPS results.

A summary table which compares the basics about [QC](#) data quality indicators on the scale of each single measurement is provided here. Again, each of the following topics is covered in much more complete detail in sections that follow, but when later reading those individual sections, readers may want to refer back to the following table when thinking through the differences between the various QC indicators, and compare their requirements with the following before finalizing QA/QC SOP details.

The following tables are relevant to chemical and biological measures. However, some ([MDLs](#) and [Blank](#) control) are most relevant to low level chemical analyses. For many biological or ecological measures, the minimum [QC](#) indicators that need to be

controlled would typically be at least three: [precision](#), [bias](#), and [sensitivity](#) (often as AMS).

QC Metric and QC Measurement Quality Indicators: Comparison of Indicator Basics

Measurement Sensitivity (Relevant to Each Single Data Point)

Typical QC control for sensitivity is done in one of the following ways explained in the table below: 1) [MDLs](#) and [MLs](#), this couplet is usually used for chemical analyses when very low-signal strength is sometimes encountered, or 2) LT-MDLs and LRLs (Only when USGS Labs are Used), or 3) [AMS](#) or [AMS+](#) (used for field measures and whenever low-signal strengths are never encountered). Each of these is summarized below:

Purpose	Metric Acronym	Metric + Brief STORET Note	Minimum Frequency of Reporting	Description	Sample Size	Equation
EPA, State, and some USGS labs Low Level Sensitivity As Detection Limits (Usually Lab)	MDL : for Control of Very Low Level Sensitivity . This is the standard MDL	Method Detection Level. Put MDL in the detection limit field. If the result is <MDL, STORET Detection Condition is "Not Detected."	1/year or when methods change	Lowest value that can be differentiated from zero, the lower Semi-Quantitative Type of Detection Limit, Based on Short Term Data	Seven	Obtain MDL from laboratory, For field work calculate as $3.134 * SD$ of a blank or very low signal solution
USGS NWQL Low Level Sensitivity As Detection Limits (Usually Lab)	LT-MDL: NWQL Control of Very Low Level Sensitivity	Long Term Method detection level USGS Long-Term Version of MDL	Every Few Years	Lowest value that can be differentiated from zero, but based on Long Term Data	High, Based on Multi-year data, Get the LT-MDL from USGS	Obtain from USGS laboratory
EPA and State Quantitative Sensitivity As Detection Limits (Usually Lab)	ML: Higher than This, Values are Quantitative	Minimum level of quantitation, In STORET, record at LQL	1/year or when methods change	Lowest Quantitative Value Above the ML values are quantitative	Based on Single MDL	$3.18 * (MDL)$
STORET synonym for ML	LQL	Lower Quantitation Limit = STORET Synonym for ML	1/year or when methods change	Lowest Quantitative Value, "	Based on Single MDL	$3.18 * (MDL)$
USGS Quantitative Sensitivity As Detection Limits (Usually Lab)	LRL	Long term reporting level (Unique to USGS laboratory)	Every few years	A USGS alternative to the ML, based on long term QC data and LT-MDLs	Based on Single LT-MDL, Get the LRL from USGS	Obtain from USGS laboratory, $2 * (LT-MDL)$
Sensitivity (Usually Field, or Whenever MDL is NA)	AMS: Lowest Change Possibly Real	Alternative Measurement Sensitivity, For \pm STORET "analytical procedure description" field	Beginning and end of field seasons	Determines instrument noise in both directions (up or down). How big of a change is real?	7 measurements from the same field sample	$3.708 * SD$, where SD = Sample Standard Deviation
AMS+ (Usually Field, or Whenever MDL is NA)	AMS+: Total Variability of Close Replicates	Alternative Measurement Sensitivity+, Record In STORET as Stated Above if no other form of AMS is reported	Beginning and end of field seasons	Includes instrument noise and natural heterogeneity	7 measurements of nearby but not identical samples, in-situ for sondes only	$SD * 3.708$

Measurement Precision (Lack of Perfect Precision, Relevant to Each Single Data Point)

At minimum, measurement [precision](#) is typically controlled either with QC duplicate measurement of a single sample for Precision or with measurement with two nearby samples for [precision+](#). These two are contrasted below:

Precision (Lab and Field)	RPD : QC Precision Control	Relative Percent Difference. In STORET include both values RPD was based on and optionally include RPD as a comment.	1 for every 20 samples, lab or field. In the field, also used for every core parameter calibration	Variability of repeated measures (precision)	1 sample but two values (1 Comparison of two values measured on one single sample)	$RPD = \left[\frac{S_1 - S_2}{(S_1 + S_2)/2} \right] \times 100$
Precision+ (Usually for Field Measurements Only)	RPD : QC Precision+ Control	Relative Percent Difference: Include in STORET as suggested above in no Other Form of Precision is Reported	1 for every 20 samples, lab or field. Done Instead of Precision or in Addition to Precision	Variability of repeated measures (precision+) + = potentially some additional true variability (two samples not one)	2 (1 Comparison of two values, measurements of two samples collected in close proximity but not one sample)	$RPD = \left[\frac{S_1 - S_2}{(S_1 + S_2)/2} \right] \times 100$

Measurement [Bias](#) (At the Scale of Each Single Data Point)

At minimum, measurement bias is typically controlled with a blind measurement of a sample when the person doing the measurement does not know the right answer when measuring. In the case of chemical lab analysis, another form of bias is controlled with [blank QC](#) samples.

Bias (Lab and Field)	PD : QC Bias Control	Percent difference: In STORET include both values PD was based on and optionally include PD as a comment, Choose Reference Sample or Field Spike	QC control 1 for every 20 samples, lab or field. In the field, also used for every core parameter calibration,	Difference Between Measured Result and Expected Result Based on a Reference Sample Standard or a Spike	2 (1 Comparison of two values, one of which is a known correct (or expected) value and the other is the measurement result	$PD = [Y - X] / X] * 100$, where X is the known (usually "correct" or "expected") or spiked amount, and Y is the measured concentration.
Blank Control Bias (Usually for Lab Measures Only)	PD : QC Blank Control Bias	Percent difference, But No Blank Contamination Positive Bias is Reported Unless The Value Measured is Higher than the MDL . Record both measured value and MDL in STORET	No Less Stringent than the State, often QC blank sample once every 20 lab samples or once per field site	Difference Between Measurement Result and Blank Sample Expected Result (Usually No Greater than the MDL	2 (1 Comparison of two values, one of which is the expected value (no greater than the MDL) and the other of which is a measurement of the blank sample.	$PD = [Y - X] / X] * 100$, where X is the MDL and Y is the measured concentration.

NPS Vital Networks may want to copy parts (or all) of the tables above into the QA/QC SOP.

Each of the above has a STORET counterpart. See table entitled "QC Measurement Quality Indicators and STORET/NPSTORET Treatment" in separate section in the last chapter herein (Include STORET Details in a Data Management SOP).

If additional detail on any of the topics listed above is located anywhere other than in the QA/QC SOP, a summary of what will be done to control each of the issues

listed above should be included in the QA/QC SOP. Point-to hyperlinks in the SOP should make it clear to the reader exactly where the other detail may be found. For example, if [representativeness](#) and [target populations](#) are fully explained in the protocol narrative or in the chapter on sampling design in the plan, then the representativeness section of the QA/QC SOP should clearly “point to” the section where the subject is fully covered. Details related to individual sites and individual measurements or parameters (“characteristics” in STORET terminology) should usually be fully explained in the representativeness section of the QA/QC SOP.

A good example of a good QA/QC SOP is the [SOP 7 \(QAPP\)](#) of the Northern Colorado Plateau Network (NCPN) attached to the [NCPN Freshwater Protocol](#).

To obtain data comparability, it is OK and even desirable to use well established QA/QC procedures of another federal agency (USGS, NOAA, EPA, NAWQA, or EMAP) or a state agency. In the QA/QC SOP, list the source-agency, measurement quality objectives, and SOP details for [sensitivity](#)/detection limits, precision, systematic error/[bias](#) and [blank](#) control (the latter for chemical labs only). The SOP source-agency (EPA, EMAP, USGS, etc.) may change their SOPs as time goes along, and we need to have solid documentation of the methods we used at the start. For subsequent method changes, (see [Include a Cumulative Bias SOP](#) section below).

In some cases, the same [QC](#) measurement quality objective ([MQO](#)) can be given for several parameters in a suite of vital signs included in one protocol. For example, if a network decided to use EPA marine EMAP QC SOPs to obtain maximum data comparability with EPA and state marine EMAP data, they could specify a precision repeatability MQO of 10% for several parameters to be measured in the field, including pH, temperature, DO, specific conductance, salinity depth, light transmittance (PAR), turbidity, and Secchi depth. The EMAP methods are a good source of MQOs for precision, bias, for field probe measures (EPA. 2001. [National Coastal Assessment Quality Assurance Project Plan 2001-2004](#). EPA/620/R-01/002).

However, in many other cases MQOs will be different for different parameters and can simply be listed in a [QC SOP table](#). A protocol for water column parameters measured in the field would typically have different measurement quality objectives than a protocol for nutrient parameters measured in the lab. However, in both cases, a table in a separate QC SOP in each protocol could list the MQOs for each applicable parameter.

Be careful with QA/QC terminology. Words and phrases such as [representativeness](#), [sensitivity](#), detection limits, accuracy, [precision](#), repeatability, reproducibility, error, systematic error/[bias](#), and uncertainty, have been used for different concepts by different groups. The confusion in water quality and contaminants QA/QC terminology has been so widespread that it brings to mind a “Tower of Babel” (everyone speaking different languages, no one understanding each other) scale of confusion. Some care has been taken herein (and in more detail in [Part B](#)) to explain the right terminology and to standardize on National Institute of Standards and Technology ([NIST](#)) and [International Organization for Standardization \(ISO\)](#) terminology wherever possible.

If the QA/QC of another agency is not adopted, or if the [QC](#) details come from multiple sources or are brand new, before completing the QA/QC SOP, a final check should be made to make sure that the measurement process will be controlled in some documented and defensible manner. The networks need to document what will be done for each of the issues listed below (doing nothing is not an option), but the networks need

not “go overboard.” However, at minimum, reviewers will be looking for common-sense documentation related to each of the following QA/QC basics:

The reason the word “measurement” proceeds each of the [QC](#) data quality indicators ([precision](#), [bias](#), [sensitivity](#)) discussed above is that what the QA/QC SOP should cover is controlling the measurement process **on the level of each single measurement**. For contrast, on the broader (**overall monitoring design**) scale of **multiple measurements**, related topics are covered in the protocol narrative rather than the QA/QC SOP. Examples of analogous but different-scale issues that would be covered in Protocol Narrative rather than the QA/QC SOP include:

Monitoring Design Sensitivity (Expressed as Minimum Detectable Differences or Minimum Detectable Effect Sizes).

Monitoring Design Precision (Which Includes Contributors to Total Variability Not only from a Lack of Perfect Measurement Precision but also From True Heterogeneity of the Variables Being Measured and a Lack of Perfect Representativeness of the Samples Measured Compared to the Underlying Target Population. This is often expressed as a standard deviation.

Monitoring Design Systematic Error/Bias (Measurement Scale Bias Plus any Bias Contributed by an Imperfect Monitoring Design Tending to Result in Consistently High or Low Estimates of the Summary Statistics (Means, Medians, Standard Deviations, Quartiles, etc.), When Compared to True Values of the Underlying Target Population). This is often expressed as percent difference.

Monitoring Design Completeness (see next section).

What will be done to control the Monitoring-Design-Scale topics discussed just may be covered in the protocol narrative rather than in a QA/QC SOP. However, summary materials may be presented in tables.

VI. Completeness, Sample Sizes, Statistics, and Detection Probabilities vs. Desired Conditions

Completeness is usually considered a [QC](#) topic, but to assure completeness, one must first consider sample sizes, overall monitoring or survey plan, detection probabilities, desired conditions and some other [QA](#) factors that are usually first mentioned in the central monitoring plan as part of the overall monitoring design. So although completeness goals are quantitative, they are different and arrived at differently than QC goals for precision, bias, and sensitivity, all of which are at the QC level of pertaining to the quality of each individual data point. Completeness, on the other hand relates to multiple data points to be collected with the overall monitoring design.

In aquatic Vital Signs Monitoring, Data completeness goals are typically given as percentages in [tables in the QA/QC SOP](#) or [QAPP](#) and are developed by first estimating required sample sizes. Although written for aquatic and water quality monitoring, a

statistician who reviewed the following section reminded us that most of these steps are generic and would also apply to terrestrial monitoring.

Determining required sample sizes and attendant completeness goals should be done in a stepwise manner, considering the following in a more detailed and quantitative way than has been done in earlier planning phases:

1. Refine (provide more time and space detail) objectives and questions
2. Identify desired conditions qualitatively.
3. Identify resource-collapse or other thresholds (such as water quality standards or no-effect levels)
4. Identify existing conditions.
5. Develop safety margin between existing conditions and threshold magnitude.
6. Document variability in time and space.
7. Refine target population details.
8. How big of a difference or change do we need to be able to detect?
9. What initial statistics will be used?
10. Choose desired detection probability/statistical power (1-beta).
11. Choose statistical significance level (alpha).
12. Use simple calculators to make initial estimates of required sample sizes.
13. Throw out measures or strata where excess variability will prevent detecting a trend or a difference of concern within budget.
14. Optimize monitoring plan details for affordability and logic.
15. Draft initial sample sizes and optimized monitoring design.
16. Finalize sample sizes and design with an applied environmental statistician.
17. Estimate the % of samples that will fail (for example 10%).
18. Increase the planned sample sizes accordingly.
19. Put completeness goals in a MQO [table](#) in the QA/QC SOP.

The first three above are typically covered in varying degrees of detail in the central monitoring plan. Some have also previously been introduced herein (above) in the section on Objectives and Questions. However, when developing the fine details of the monitoring design, sample sizes, and statistics, several of these inter-related issues should be reconsidered in a more thorough and quantitative way and documented in more detail in each protocol narrative and in relevant SOPs. The goal would be to make sure they all line up and make sense when considered together. Defining the first two steps in as much time and space detail as possible is helpful when moving to the more quantitative steps (3-19).

When faced with a 19 step process (just above), why not just go to a professional statistician to start with (or maybe starting along about step 4)? Great idea, if the network can afford it. However, many of the steps are decisions to be made by the park or network, not the statistician, and would in fact be input to bring to the statistician. All of the steps before 16, except perhaps 9 and 12, should be done by NPS staff, often with the help of the network quantitative ecologists. Even if performed by a statistician, the statistician would need considerable input from the NPS in going through the steps. Furthermore, bringing Vital Sign network quantitative ecologists up to a certain minimum level of understanding is a good goal and one that would help prevent some

past disconnects between the statistician's advice (often in Chapter 4 of the central monitoring plan and mistakes made later by networks in protocol and SOP development after the monitoring staff stopped talking to the statisticians.

Fully informed quantitative ecologists can help park management refine the steps above (in developing sample sizes, minimum detectable difference goals, power goals, etc. in an adaptive management way (see [Statistical Methods for Adaptive Management Studies](#)). For example, after step 15 above, it may become clear to all that initial decisions made for steps 1, 5, 6, 8, and 11 have to be adjusted for the design to make sense and be within budget.

In situations where: organisms are clustered or clusters are geographically rare; moving, or are potentially newly establishing invasive species, then adaptive sampling can be done to find and then more intensively sample locations where rare species live or rare conditions can be found. This is a different topic. Sample sizes needed, and how to estimate variance tends to be done in different ways. Adaptive sampling issues can be complex (USGS 2002. [Sampling Designs for Rare Species and Populations](#)). Software Downloads for [Adaptive Sampling](#) are available from USGS.

There are also cases where parks or networks might want to monitor rare charismatic species or rare resources with legal mandates. However, in many other cases rare resources are excluded from long term monitoring for various practical reasons. Among the factors to consider: 1) the rare resource might disappear during 100 year monitoring, 2) the resource might be impacted by monitoring repeatedly, 3) logistical and other costs can be high, especially when trying to sample rare resources of unknown distributions using multiple-step adaptive sampling, and 4) often there is not enough power to make any kind of statistically powerful claim about the resource in decline until the resource is already extirpated or otherwise in a difficult-to-recover situation.

No matter what type of sampling is planned, determining required sample sizes and data completeness goals admittedly takes a bit of effort. However, is especially important for long term monitoring and failing to carefully plan for adequate sample sizes has all too frequently resulted in aquatic monitoring that has produced data that has not been useful for management decisions. Too often raw data has never made the transition to useful information ("Ward, R.C., Loftis, J.C., and G.B. McBride. 1986. The "data rich but information poor" syndrome in water quality monitoring. *Environmental Management* 10:291-297).

Outliers:

If sample size is large enough (usually 25 or more), one way to identify "potentially wrong" outliers is to calculate the 5th and 95th percentiles (some use 10th and 90th percentiles) and then look closer at those values that fall outside of the those limits. Outliers tend to strongly influence means and standard deviations.

If the outliers seem clearly wrong (like impossible pH values of 85 or -100), they can be discarded, but be careful not to discard values without strong evidence that they are clearly wrong, since often extreme values can be right and in fact the most important ones (Helsel and Hirsch text

book (Helsel, D.R. and R.M. Hirsch. 2002. [Statistical Methods in Water Resources](#). US Geological Survey Techniques of Water Resources Investigations, Book 4, Chapter A3). See also [part B](#) for more details on outlier STORET codes.

A bit more detail on each of the sample size outline steps is provided as follows:

1) Refine (Provide More Time and Space Detail) Objectives and Questions

Why revisit and refine questions? The monitoring design and statistics to be used are both driven by the questions to be answered, and it helps if the questions to be answered (and the identified target population being monitored) are as detailed in time and space specifications as possible (see earlier sections on Questions and Objectives and on [representativeness](#) and [Target Population](#)). Again, it is very important that all of the concepts in the following outline line up and be reasonable when all are considered together.

If one calls them objectives rather than questions, the details of what, where, when, and (even) how big of a change can we detect; all still need to be detailed before one can design monitoring in an optimal way. The number of objectives competes with the number of samples in a cost-limited study (Kurt Jenkins, USGS BRD and North Coast and Cascades Network, Personal Communication, 2006).

2) Identify Desired Conditions Qualitatively First

The central monitoring plan should document desired future conditions (more recently referred to in the NPS as desired conditions or DCs). At the protocol development stage, additional detail on targets for individual water quality parameters or other indicators to achieve DCs should be placed in the protocol narratives. For a generic (not just water) Vital Signs monitoring discussion of DCs, see [talks by Steve Fancy and Rob Bennetts](#) at the San Diego NPS Vital Signs Meeting in 2006. Some of the key points therein and other related points are summarized briefly as follows:

DCs are general qualitative descriptions used at the General Management Plan stage. In the new-style NPS General Management Plans, DCs are defined as “A **qualitative** description of the integrity and character for a set of resources and values that park management has committed to achieve and maintain” (NPS, 2005. [Park Planning Source Book](#) and [Appendices and glossary portions of the revised NPS Planners Sourcebook](#), both available on NPS computers only).

Additional insight may come from the new-style Resource Stewardship Strategies (RSSs, to be completed after the revised General Management Plans). At the later RSS stage, NPS staff members typically attempt to become more quantitative with goals (individual quantitative targets) for each of multiple water quality indicators. Together these multiple quantitative goals should help insure the much more general and qualitative desired condition statements.

Recently completed NPS Watershed Assessments are good sources of information related to current conditions versus targets needed to achieve desired condition goals, for

those Parks that have had them done. The Watershed Assessment Program hopes to complete all 278 Natural Resource (I&M) Parks by 2014. For details, see NPS intranet-only homepage (NPS. 2007. [Program Plan to Assess Watershed Conditions](#)).

When considering reference conditions and targets for desired future conditions, monitoring networks should be sure to consult Park Superintendents and Park Resource Management Staff for their thoughts on DCs (and potential quantitative targets to help achieve DCs) from a management perspective.

The final 2006 NPS [Management Policies](#) document notably emphasizes both protection of park resources for the enjoyment of future generations and sustainability of both natural and financial resources. With certain exceptions (big recreational use reservoirs, and a few other scenarios), the policy also notably emphasizes protection of native species. However, in recognition that conditions are not static, Section 4.4.2 addresses the fact that natural processes will be relied upon to maintain native plant and animal populations.

Section 4.4.4.2 of the NPS [Management Policies](#) 2006 document specifies the following for Removal of Exotic Species Already Present:

All exotic plant and animal species that are not maintained to meet an identified park purpose will be managed—up to and including eradication—if (1) control is prudent and feasible, and (2) the exotic species interferes with natural processes and the perpetuation of natural features, native species or natural habitats, or disrupts the genetic integrity of native species, or disrupts the accurate presentation of a cultural landscape, or damages cultural resources, or significantly hampers the management of park or adjacent lands, or poses a public health hazard as advised by the U.S. Public Health Service (which includes the Centers for Disease Control and the NPS public health program), or creates a hazard to public safety. High priority will be given to managing exotic species that have, or potentially could have, a substantial impact on park resources, and that can reasonably be expected to be successfully controlled. Lower priority will be given to exotic species that have almost no impact on park resources or that probably cannot be successfully controlled. Where an exotic species cannot be successfully eliminated, managers will seek to contain the exotic species to prevent further spread or resource damage.

Section 4.4.4.1 outlines circumstances under which exotic species may be maintained (generally only to maintain cultural landscapes or resource condition (i.e. non-natural parks or areas of parks), or to control other already established exotics.

In other words, desired conditions, or even possible conditions, tend to be moving targets (and to vary within fairly broad ranges) due to purely natural processes. There are also anthropogenic stressors of various types, including invasive species, various direct and indirect human impacts, and countless other changes in stressors upon which NPS resource managers often have limited control.

Related emerging concepts and complications include applied historical ecology, the fragmentary nature of history, the subjective and value-laden aspects of desired condition goals, and pre-Columbian impacts of man. Other issues include assumption, difficult-to-quantify confounding factors, no-modern-analog issues, and non-equilibrium

paradigms. Thus it is sometimes fruitless to choose a single fixed point goal and better to use ranges [Swetham, T.W., C.D. Allen, and J.L. Betancourt, 1999. [Applied historical ecology: using the past to manage for the future](#). Ecological Applications 9(4)].

How would global climate change influence NPS management and monitoring goals? Section 4.7.2 of the final 2006 NPS [Management Policies](#) states that

“Earth’s climate has changed throughout history. Although national parks are intended to be naturally evolving places that conserve our natural and cultural heritage for generations to come, accelerated climate change may significantly alter park ecosystems. Thus, parks containing significant natural resources will gather and maintain baseline climatological data for reference.”

However, the above paragraph still leads conscientious scientists and resource managers to contrasting opinions. Part of the debate relates to the fact that some tend take a longer term look at the issue than others. An interesting expanded discussion of “what is natural” and conserving biodiversity (including rare species, and how native species are often difficult to list based on our short term records, and dynamics of species movements) is found in Willis and Burks, 2006 (K. J. Willis and H. J. B. Birks 2006. What is natural? [The need for a long-term perspective in biodiversity conservation](#). Science 314:1261-1265).

If a non-native species eventually become “naturalized”, are not causing problems for native species, are helping rather than hurting for biodiversity, are appropriate due to changing climates, and/or are serving a vital ecological function that might otherwise be missing, etc.), NPS unit managers may eventually decide that “observed to desired” (O/D) species ratios are more useful than strict observed to expected (O/E) species ratios related to native taxa loss only.

In other words, over the long run, not every scenario will be optimally covered with a strict (O/E) only. As mentioned earlier, there are some exceptions to general NPS [Management Policies 2006](#) related to how recreation reservoirs are to be managed. In such habitats, NPS Recreation Area Managers may have the flexibility to choose not to try to get rid of non-native lake-species of fish simply because they are not native. In many cases, the natives were riverine fish, so their native habitats have been altered. In such cases, observed to desired (O/D) species ratios may be helpful. Some of the large NPS managed reservoirs have fishing recreation recognized in their enabling legislation.

As can be seen in the discussion just above, these topics are complex and should be handled with some care, and only approached with involvement of NPS unit resource managers and ideally with input from Superintendents.

Some States (albeit not without some controversy) have written Water Quality Standards language in a way that does not ignore existing realities by describing State reference conditions as “best available representatives of ecoregion waters in a natural condition.” EPA’s guidance on reference conditions for bioassessment likewise acknowledges that “Recognizing that pristine habitats are rare (even remote lakes and streams are subject to atmospheric deposition), resource managers must decide on an acceptable level of disturbance to represent an achievable or existing reference condition” (EPA. 2007. [Aquatic Life Use Support \(ALUS\)](#)).

In a similar manner, some NPS-unit managers may decide it is OK to set short-term goals for desired future conditions that are realistic considering foreseeable realities of surrounding areas. Longer term, these same managers may wish to consider desired future condition goals that more closely reflect more minimal anthropogenic effects (to the degree possible).

Moving From Qualitative To Quantitative Goals

As introduced above, the step of developing a new-style Resource Stewardship Strategies (RSS) is typically where the NPS starts making general DC goals more quantitative, by translating the general goals to more specific quantitative targets for individual water quality parameters. However, even if the park has not yet completed a quantitative RSS yet (and few had been completed by January, 2008), monitoring planners need to have some quantitative targets in mind for individual water quality parameters and indicators. Even if such targets are just the result of an initial best professional judgment estimate, planners need quantitative targets before they will be able to compare existing condition to targets. That information is needed before monitoring planners can then compare calculated minimum detectable differences with ideal magnitudes of differences that monitoring needs to be able to detect. In other words, regardless of where the Park is in developing quantitative RSSs, monitoring planners need quantitative goals. How does the network propose doing this? Here is one recommended thought process.

Consider NPS Impairment Guidance

It is typically appropriate to think through where we are now and how big of a change from the current condition would it take to move beyond a negligible, minor, or moderate impact as defined in NEPA terminology (NPS 2003 [Interim Technical Guidance on Impairment of Natural Resources](#) (available on NPS computers only). Therein, parks decide whether or not a **biological** impact is negligible according to the following definition):

Negligible Biological Impacts: Impacts occur, but are so minute that they have no observable effects on plants and animals and the ecosystems supporting them. The severity is “Trivial effects on individual organisms or areas of habitat.” The duration is “Short-term to long-term effects.” The timing is: “Outside of critical timing windows of key resources or ecosystems.”

Parks decide how big water quality impacts are according to the following criteria):

Negligible Water Quality Impacts are described as “Impacts are effects that are not detectable, well below water quality standards, and within historical baseline water quality conditions.” The impairment guidance (op. cit.) also describes other levels of water quality impacts:

Minor: Impacts are effects that are detectable but well within or below water quality standards and within historical baseline water quality conditions.

Moderate: For most waters, impacts are effects that are detectable, within or below water quality standards, but historical baseline water quality conditions are being altered on a short-term basis. However, in outstanding natural resource waters (ONRWs), this threshold may approach the requirements for statutory impairment.

Major: For most waters, impacts are effects that are detectable and significantly and persistently alter historical baseline water quality conditions. Water quality standards are locally approached, equaled, or slightly and singularly exceeded on a short-term and temporary basis. However, in ONRWs this threshold would probably constitute statutory impairment.

In highly pristine resource parks or parks with highly valued, rare, or endangered resources will DCs be defined as negligible impacts according to the above?

If the water is already much cleaner than default state water quality standards, will stronger quantitative water quality goals or anti-degradation standards be used?

At the other end of the spectrum, if one is considering a historical park in a highly urbanized, industrialized, or farmed area where there is no reasonable expectation of ever achieving negligible impacts, the area may be in an alternate steady state. In this scenario, targets related to desired future conditions (DCs) may simply be avoiding additional significant impacts (anti-degradation). Another goal might be improvements in condition when the chance arises, or perhaps meeting urban habitat state water quality standards.

Consider O/E Goals

Observed to expected ratios (O/E) can be helpful when trying to develop quantitative desired condition goals. Such ratios are sometimes used in conjunction with predictive models to contrast existing conditions to specific targets related to more general desired conditions (DCs). O/E ratios and models have the advantage of being so easily understood that they are intuitively appealing.

Recently the EPA 2006 nationwide [Wadeable Streams Assessment \(WSA\)](#) prominently used O/E primarily in the context of **native benthic macroinvertebrate taxa lost**. The WSA noted that under the Clean Water Act, non-native species that compete with and potentially exclude native species might be considered simply another stressor and in fact a threat to biological integrity. In the WSA, the Macroinvertebrate O/E Ratio of Taxa Loss measures a specific aspect of biological health: taxa that have been lost at a site, so non-native species are presumably not counted in either O or E metrics. The taxa expected (E) at individual sites are predicted from a model developed from data collected at least-disturbed reference sites; thus, the model allows a precise matching of sampled taxa with those that should occur under specific, natural environmental conditions. By comparing the list of taxa observed (O) at a site with those expected to occur, the proportion of expected taxa that have been lost can be quantified as the ratio of O/E (page 31, [WSA](#)).

The decision to include non-native but otherwise desirable species in either O or E metrics is a policy issue, not a scientific issue. Even in O/E ratios, non-natives could be included if they were considered to be a valuable resource (Chuck Hawkins, Utah State University, Personal Communication, 2007).

To avoid terminology confusion, it might be better to call such ratios something else. . In some scenarios the O/E ratio could then become something closer to a “valued species observed to total number of species at the site under desired conditions” (or observed to desired --O/D for short) ratio or something similar rather than a more traditional O/E ratio based on native-species only. Therefore, some investigators might be flexible on including certain desired or naturalized species (presumably those not overwhelming or otherwise causing significant problems for native species, ecological functions, or total biodiversity) on either the O or the E parts of the ratio. Large NPS reservoirs having recreational fishing prominent in their enabling legislation, where non-native fish are stocked on purpose, constitute one exception to “native-only” policies. Such exceptions are specifically mentioned in NPS [Management Policies](#) 2006.

However, except for these types of named-exceptions, many are understandably reluctant to move away from native-species-only goals, partly due to established laws and policies.

For example, the protection of native species is not only emphasized by NPS policies, but also language of the Clean Water Act (CWA). Therefore, “when non-native species become established in either vertebrate or invertebrate assemblages, their presence conflicts with the definition of biological integrity that the CWA is designed to protect (i.e., having a species composition, diversity, and functional organization comparable to that of the natural habitat of the region).” (EPA. 2006, [WSA](#)). O/E ratios for benthic macroinvertebrates in wadeable streams of the U.S. are being improved to the point of becoming more and more broadly usable (see [Predictive Models Primer](#) of Utah State). Utah State, in conjunction with EPA EMAP and various state and federal agencies has worked out regional models for the Western US, Eastern Highlands, and the Midwest (EPA. 2006. [Wadeable Streams Assessment](#)).

These regional models work fairly well for large regional scales but not (yet) always optimally well for site-specific condition assessments, where sample size may still be too low and/or local conditions may require a more site- or region-specific model. Utah State has developed state-specific models for OR, WA, CA (3), WY, MT, and CO. Models for ID, UT, and AZ will be completed soon. Not always fully understood is that it is really more accurate to talk about 'sample' O and E and not 'site' O and E. It is not yet clear how well sample O/E scales to true site O/E. The same is true for other metrics too (Chuck Hawkins, Utah State University, Personal Communication, 2006).

Northern Cascades National Park staff members are developing a network-specific O/E predictive model (similar to the one used by the EPA [WSA](#)) to use at North Coast and Cascade network Parks (Reed Glesne, NPS, Personal Communication, 2006) and similar efforts are underway for Rocky Mountain Network Parks (Billy Schweiger, NPS, Personal Communication, 2006).

Until more park, region, state, or network-specific O/E predictive models are available, state water quality standards, and comparison values from state or regional multi-metric biotic integrity methods are often used as interim quantitative comparison benchmarks. If the water is already much cleaner than default state water quality

standards, Parks and networks often need to decide whether or not stronger quantitative water quality goals or anti-degradation standards will be used to estimate desired future conditions.

Even when highly refined O/E predictive models are available, many believe that we should also look at other lines of evidence, such as evidence from multi-metric methods, multivariate methods, chemical evidence, and physical habitat evidence. Although it is one of the strongest lines of evidence, O/E tends to focus on only one line of evidence (native taxa loss).

EPA's 2006 national Wadeable Streams Assessment ([WSA](#)) found that O/E results tended to track but not exactly replicate multi-metric methods that looked at more lines of evidence. In its executive summary, EPA summarized the nationwide stream condition results from different regions based on its "Macroinvertebrate Index of Biotic Condition" rather than O/E, which was first mentioned in the context as supporting evidence from a single (albeit very important) line of evidence on page 39 of the report. So although O/E has particular appeal, we should probably continue to examine relevant data from all different angles including all available multiple lines of evidence.

In addition to O/E, other intuitively appealing and easily understandable "report card" type metrics that cross different types of resources (aquatic, terrestrial) etc. (and are potentially useful for GPRA type reporting goals) also often use proportions. Among these are the % of time not meeting standard (or exceeding a threshold), % of river miles impaired, % of lake or ponds (or their surface area) impaired, etc.

Iterative Goal Setting

Initial quantitative goals to help achieve desired conditions (DCs) need not be permanent. Gradually refining DC goals in a classic DOI-style adaptive management cycle might follow a pattern such as this (expanded slightly but based on suggestions by S. Fancy. 2006. [Desired Future Conditions](#)).

1. Park staff qualitatively describe the DC in the new General Management Plans,
2. Park staff compare current conditions for various water quality parameters and other indicators to targets needed to help achieve DCs. Decisions should be documented in the new cycle of developing Resource Stewardship Strategies (RSSs) that include quantitative performance measures and goals. This might typically include listing current % of river miles impaired vs. goals in the same units relevant to GPRA and DCs.
3. Park staff members develop and implement management strategies to achieve desired conditions.
4. Park staff (at times possibly supplemented by Vital Signs planning and monitoring information already completed) finishes developing quantitative performance measures for monitoring. At this stage it would be optimal to be as quantitative as possible and define minimum detectable differences of the monitoring design.
5. Monitoring is done to detect trends in resource conditions and evaluate management effectiveness (Sit, V. and B. Taylor (editors) 1998 [Statistical](#)

[Methods for Adaptive Management Studies](#), B.C. Min. For., Res. Br., Victoria, BC, Land Manage. Handbook. No. 42.).

6. Park management uses the results from monitoring to take adaptive management actions to help achieve desired conditions,
7. Park staff returns to (1) above, and they refine quantitative targets to help achieve DCs, if appropriate, and
8. Park and monitoring staff keeps repeating the cycle and refining each step as appropriate as lessons are learned.

3) Identify Resource-Collapse and Other Thresholds of Concern

In the absence of other more detailed information on thresholds, one often uses water quality standards that already have a safety margin built-in.

Park resource managers would typically desire to detect a change smaller than one that would change conditions from what they are currently to a condition that no longer meets water quality standards. They also would typically want to detect changes smaller than a change that might move conditions across a threshold that would cause a resource collapse.

The monitoring protocol narratives should identify a resource collapse threshold for each vital sign or measure, if one is known, but often resource collapse thresholds are not known. However, one advantage of long term monitoring is that our understanding of resource dynamics and thresholds and threshold models can be refined in an adaptive management fashion as more data is collected.

There are many examples of this in fishery literature. A resource collapse in one location sometimes identifies the threshold and the threshold can then be used to protect against collapses in other similar habitats in the region. Commercial fish interests and some fishery regulatory agencies might manage a fishery to sustain less than 50% of virgin biomass (sometimes goals are as low as 30%) to strike a political/societal balance between sustainable fishery yields and economic benefit. Of course, this is a strategy that could radically change species relationships away from normal conditions. In the North Atlantic Cod fishery, not only was the natural biomass of Cod greatly reduced, but shellfish numbers and catches greatly increased, partly due to the reduction of their predators (Ray Hilborn et.al. 2003, [State of the World's Fisheries](#), Annu. Rev. Environ. Resour. 28:35).

For contrast, in the NPS we would ordinarily try to manage species assemblages and biomasses to reflect more natural conditions (far more than 50%, probably often closer to 85% plus or minus 15% or something similar for virgin biomass). The NPS would also tend to manage for normal biotic assemblages.

Sometimes there are no resource collapse thresholds available and one has to use other quantitative comparison benchmarks. EPA Costal EMAP has suggested regional coastal threshold criteria for many indicators, including chlorophyll a, benthic indices, water clarity, and DIN and DIP (EPA et al., 2007, [National Estuary Program Coastal Condition Report](#), Chapter 2).

Various water, sediment, tissue, and soil benchmarks are also often used. Risk assessment-derived benchmarks, especially No Observable Adverse Effect Levels, No Effect Concentrations (NOAELs, NOECs), and Low Effect Levels, can be used as one

starting point, as long as it is realized that they are often less than optimal since they are seldom based on considerable local or regional work or even sound statistics. However, in some cases one has to use what is available until something better and more defensible is available (adaptive management approach).

Summaries on data comparison benchmarks for metals and industrial organics and petroleum hydrocarbons in water, sediment, soil, and tissues, updated through 1997-1998, are summarized in the [NPS Contaminants Encyclopedia](#). The NPS encyclopedia contains general ecological toxicity profile information on 118 contaminants ([listing of all 118 topics](#), and for [references for all 118 topics](#)).

Similar [hazard profiles for human health](#) instead of ecological resources are available from ATSDR.

Whether or not there are state water quality standards, it is often desirable to compare local conditions to regional benchmarks and/or the most pristine areas in the region.

Change point analyses are data-hungry and computationally complex but are sometimes used as one line of evidence for thresholds. However, such analyses are no more definitive as a stand-alone evidence of causation than correlation analyses (see N-steps [change point analysis introduction](#) and [correlation introduction](#). Change point analyses are usually used in conjunction with regression techniques to plot and otherwise analyze the relationship between two variables (see [regression introduction](#)).

.Thresholds, including water quality standards, should be thought of in the context of “if-then” management decision rules. If damage or toxic concentrations exceed such and such and magnitude, then we will do what? Will we reduce visitation, begin remediation, or reduce fishing pressure (or what, see section IV-C of [Part B](#))?

4) Identify Existing Conditions

This usually includes the previously discussed analysis of [past data](#). In some cases, pilot-scale studies will have to be done when little or no information is available.

5) Develop Safety Margin between Existing Conditions and Threshold Magnitude

How much does the ambient condition have to change to get to a threshold of concern or a specified desired future condition? If a resource collapse threshold is known, how much does the ambient condition have to change to get the threshold (the value below which the resource will not recover, or will recover at an unacceptably slow rate, also called a breakpoint by some)?

Again, if one is only 10% of the mean away from disaster, then obviously being able to detect an effect-size change of a magnitude of 20% of the mean is not good enough. Water quality standards usually have a safety margin built in, but many other comparison benchmarks do not.

Too often, monitoring projects fail to detect true anthropogenic effects (type II error, false negatives, the conservationist’s risk) because of inadequate survey design. In studies that measure change, there must be a large enough sample size to detect the minimum effect, or smallest difference or change that will cause management action. The

smallest change is usually defined in terms of an “effect size” or minimum detectable difference ([MDD](#)).

How is the effect size expressed? Is it in original units as a difference between a mean and a water quality standard or the difference in two means? This is a common definition used by and is also used by Zar (op.cit.) and a few others.

Or is “effect size” the totally different concept used in behavioral sciences and recent water quality work, a percent of the standard deviation (difference divided by the standard deviation all times 100 to get a percentage, where the standard deviation is the pooled within groups standard deviation)?

Or is effect size a percentage change in a proportion, or a model output function?

In the case of fisheries, thresholds have usually been defined following collapses, trial and error, and (sometimes) long recoveries. One reason for being precautionary is that some collapses can be permanent, with recovery never occurring. If a species is on the edge of its range, or just making it for whatever reason, even a minor change, such as a climate change or new competing species, might prevent recovery. Or, as is the case for some endangered species, populations can sometimes simply become so small that they don’t survive.

In one example that considered both collapse threshold (magnitude) and smaller (safety margin) effect sizes to be detected, a slow growing macro algae, *Hormosira banksii*, was found to readily recover from depletion down to 30% cover. Pilot studies indicated an average cover of 75-85% cover. To give some margin of safety, the critical effect size goal was determined; monitoring needed to be able to detect a 30% or greater reduction in cover. This would allow detection of a reduction of cover from 75 to 45% and would also provide an effect size a safety margin of at least 1.5 (45/30) times compared to the collapse threshold of 30% cover. The idea was to give management time to institute protective strategies well before the threshold of 30% total cover might be reached (Mapstone, B.D. 1995. Scalable decision rules for environmental impact studies: effect size, Type I and Type II errors. *Ecological Applications* 5: 401–410).

6) Document Variability in Time and Space

The protocol narrative should document what is known about variability patterns in both time and space. Whenever practicable, target population definitions and sample sizes estimates should not be finalized without the best available estimates of variability patterns. Typically, the better such estimates are, the better one can design monitoring in an optimal way.

Estimates of variation can vary according to sample size and according to the magnitude of signals and means. Exclude variability estimates if they are mostly based on signals not more than twice the [MDL](#) low-level detection limits, since variability in measures that close to the very lowest detection limits would typically be much higher than for normal measurements ([Part B](#)).

Calculating the initial probability of detecting an effect size (as a % of the standard deviation) does not require input of a variance or nonparametric analogs. However, as soon as such estimates are available, they should be documented so that they can be used as input variables in estimating a minimum detectable difference ([MDD](#)) in original units of measure.

It is best to also have MDDs in original units of measure, since they are more intuitive and understandable as well as typically being more comparable with water quality standards and other comparison benchmarks. It is also optimal if the initial variability magnitude estimates come from the areas to be sampled. If no such information is available, past data from similar habitats in the region will often be an acceptable starting place for initial estimates of variability. If there are no nearby data from similar habitats from past monitoring, pilot-scale monitoring may be needed.

Most of the required-sample-size calculations related to comparisons with means depend on a good estimate of the standard deviation (SD), so be wary if the SD estimate is based on sample sizes under 30-50 unless variation is known (for sure) to be VERY small. Also be wary if the values used to estimate a Standard Deviation do not cover the full range of time and space conditions of the identified target population. If the variability is very low and the full range of conditions has been covered, a few samples may be enough, but a few samples may be virtually worthless in the presence of substantial variation (a common case for most water column measures). The same is true of most other summary statistics (not just SDs)

Low sample sizes can not only be a problem related to hypothesis tests, but also can decrease the accuracy with which we can estimate a confidence interval about a mean or proportion.

In the case of proportions, the caution about small samples (less than 30-50) also applies, see EPA discussion of [“why a sample size of 50”](#). Estimating needed sampler sizes for proportions is in some ways simpler, since the sample SD is not an input for sample size calculators. If we compute a standard error of a proportion (p) from a sample, the standard error of that proportion would be estimated as follows:

$$\sqrt{\frac{p \times (1 - p)}{n}}$$

Thus, the larger the sample size (n), the smaller the standard error (or other confidence interval) about the estimated proportion. Statisticians refer to this as “precision” (sic), though it is really about the magnitude of a half width of a confidence interval ([the part on the either side of the mean](#)) from multiple data points rather than the more familiar [precision in QA/QC settings](#) relevant to each data point. See USGS discussion for a typical statistical-specific special definition of statistical precision as a [confidence interval half-width](#).

In the case of the standard error of the mean (SEM), by contrast, a typical estimate the magnitude the SEM is dependent on both sample size and the sample standard deviation (SD).

At very small sample sizes the sample SD may over or underestimate the true population SD, due to the particulars of the specific sample, but in general, as Zar (op.cit.) explained, the sample SD tends to slightly underestimate the true but unknown population SDs, less so at large sample sizes. Other generalizations about SDs vs. sample sizes are typically problematic and potentially misleading.

Confidence intervals are about uncertainty rather than just about variability, and a key to understanding the relationships between SEMs (or other varieties of parametric

CIs about means) and standard deviations is that (if sample size is large enough) one can have high data variability (i.e., a “large: standard deviation) while at the same time having a “small” standard error about the mean. That is, even if data are very variable, if one has a large enough sample size; one will have a very small confidence interval about the mean, reflecting a comparatively accurate estimate of the mean. For more details, see page 50 of McBride’s statistical text book (McBride, G.B. 2005. [Using Statistical Methods for Water Quality Management: Issues](#), Wiley, NY, 313 pp.).

If one understands variability in time and space well enough (and statisticians will point out that this is often not true), stratified random samples, sometimes accompanied by narrow temporal collection index-period windows, can be used to reduce variability and make it easier to detect trends between years or over a period of several years.

For contrast, simple random sampling or sampling at various times of year (or even various times of day) can 1) increase variability greatly for many water column or sediment quality parameters, and/or 2) greatly increase the number of samples needed, and/or 3) decrease our ability to pick out a signal (true change of a certain magnitude) from the background of natural “noise” variability, and/or 4) decrease our confidence in the magnitude of the signal we will be able to detect), and/or 5) increase cost, and 6) often results in clumped samples that are not spatially balanced. Also, considerable seasonal, temporal, or microhabitat-type driven variability is more common for contaminants sampling in water or sediments than has been widely recognized.

Again, too often strata or index periods are picked based on untested assumptions about patterns of variability. The more one understands variability in time and space, the better job one can do of making decisions about potential strata and index period windows of time (and/or space) to sample.

For example, in diatom work done in Idaho streams, sampling date had a much stronger effect on assemblage compositions than sampling year or the sampling location within a reach, a pattern one might not have intuitively guessed before sampling began (Cao et al. 2006, [Sources of Error in Developing Biotic Indicators for Diatom Assemblages in Idaho Streams](#), NABS Anchorage Mtg. abstract).

General monitoring design theory holds that "if the response of interest displays substantial variation in one aspect of time or space, but not the other, we need to sample across the variable dimension, but can more or less ignore the other with little loss of information" (Scott Urquhart, Department of Statistics, CSU, Personal Communication, 2006).

How large does variation from one component have to be before it becomes dominant over a smaller component? One simplified rule of thumb is that if one standard deviation is five times larger than another, than the larger one becomes dominant and the smaller one becomes relatively insignificant (United Kingdom Accreditation Service, 2000. [The Expression of Uncertainty in Testing](#), UKAS Publication ref: LAB 12).

In the water quality world, it is often handy to assess which variance contributors are driving most of the variability. However, a typical problem we have for **water column measures** is that such measures tend to vary not only by space, but also by time of year and even by time of day. So all readily apparent dimensions (including residual variation from lack of perfect measurement [precision](#)) appear to contribute to variation, and there appear be none that we can automatically ignore.

This may present an example where the only way to find a year to year trend would be to use measuring instruments that can measure very precisely (to minimize residual variation) and to narrowly limit index collection time and space-periods in more than one dimension (to try to get true variability down). For example, a protocol might call for sampling only during mid summer low-flow AND only in the morning, and to sample only in full-mixed and/or narrowly-defined microhabitat strata. Again, when making these kinds of decisions, the more we understand about variation in time and space (and from lack of perfect measuring instrument precision), the better off we are.

7) Revisit and Refine Target Population Details

Previous sections introduced [Representativeness](#) and [Target Populations](#). When one is refining protocol details by moving into calculating required sample sizes, it is a good time to revisit what will be done to assure representativeness and how large target populations should be given funding limitations.

Target populations should be identified as narrowly as possible in time and space. Make sure that identified target populations line up with questions, proposed monitoring design, and other factors in this outline. For example, does the monitoring design ensure that the samples will be fully representative of the full range of values in the target population, considering what is known about variability in time and space? Is the target population all values that could be measured in bluegill sunfish in the park, or daytime bluegill sunfish between length A and B, only in limited and defined-size-range of small ponds that are road or trail accessible?

8) How Big of a Difference or Change Do We Need to Be Able to Detect?

What magnitude of change (or difference vs. a water quality standard or other comparison-benchmark) do we need to be able to detect? Once initial qualitative decisions about desired conditions (DCs) have been made, to intelligently design long term monitoring and to decide and document the magnitude of minimum required-sample-size targets, there is usually no getting around the tough decision of what is the QUANTITATIVE minimum detectable difference ([MDD](#)) that we need to be able to detect. Sometimes the MDD is expressed as an “effect size” (hereafter, abbreviated to ES). Whether we call it a MDD or an ES, it is the magnitude of change that we need to be able to detect in order to be able to manage the resource in an optimal, protective, and precautionary manner. This is a decision to be made by NPS staff, not statisticians.

Monitoring Design Sensitivity vs. Measurement Sensitivity

Just as there are detection limits (such as method detection limits/[MDLs](#), see section farther below on QC measurement sensitivity) for single measurements, on a higher level of organization (multiple measurements), a given monitoring design will have a detection sensitivity (minimum detectable difference, [MDD](#)). The MDD magnitude is driven by variability, sample sizes, significance level selected, power magnitude selected, and various other details.

No matter the scale or level of organization, sensitivity always relates to signal to noise ratios and how small of a difference we can detect quantitatively.

Lack of perfect measurement [precision](#) contributes to variability magnitudes (on the scale of each data point). True heterogeneity contributes variability (on the monitoring design scale). Both contribute to the overall variability of the values recorded.

How would one compare the magnitude of these two contributors to total variability? Lack of perfect measurement precision simply adds variability above and beyond true heterogeneity, so it is already factored in when samples from the [Target Population](#) are measured. On the other hand, if sample sizes are too small and/or the full range of the target population are not fully covered by the sampling scheme, true variability may be estimated poorly (often underestimated). This is yet another contributor to poor estimates of true heterogeneity. Assuring [representativeness](#) is crucial. Herein, for clarity, we do not use the phrase “sampling error” when the concept being discussed is really just variability rather than systematic error/bias.

Assuming representativeness is well assured, frequently the next question of interest then becomes whether or not measurement uncertainty is so large that it is significantly impacting our estimates of true heterogeneity in variables.

One can combine sources of uncertainty in sum of squares equations for either [combined or expanded uncertainty](#).

As mentioned before ([Why Document QC](#)), no (single) measurement is perfect. Each is an approximation and individual measurement data points are not complete unless accompanied by a statement about the uncertainty of that approximation.

So ideally, one would not report a single measurement of say dissolved oxygen as 5.0 but rather 5.0 ± 0.2 if 0.2 was the calculated [NIST/ISO](#) expanded uncertainty. In water quality monitoring, this has not yet gained wide acceptance, but it should be done more in the future, especially if properly-framed confidence intervals about summary statistics (means, median, SDs, etc.) are not calculated and provided to data users along with sample sizes. If the summary statistics are given with confidence intervals, then the increased variability brought on by single-measurement uncertainty is already factored in and the single measurement intervals would just give one an estimate of the contribution of measurement uncertainty versus other contributors to total variability.

One handy rule of thumb used in the United Kingdom (UK) is that (in root sum of square equations), there is no need to add negligible variance [each square of a standard deviation (SD) is a variance] terms. Although adding them would be the safest (especially if there are many such terms to complicate the issue) in many cases if any of the standard deviations is so small that their contribution to overall uncertainty is negligible (the standard deviation is at least five times smaller the standard deviation of the next largest contributor to uncertainty), they may often be ignored. In other words, regarding “domination of the combined value” (in sum of squares equations) by one component (expressed as a standard deviation), “there is not a clear-cut definition of a dominant component but a practical guide would be where one component was more than five times greater than any other” (United Kingdom Accreditation Service, 2000. [The Expression of Uncertainty in Testing](#), UKAS Publication ref: LAB 12.). In some further clarifications that include Bayesian and other more complex discussions, UKAS continues emphasizing the importance of dominance, giving separate ways to handle parametric and non parametric cases: “The exceptional case arises when one contribution

to the total uncertainty dominates; in this circumstance the resulting distribution departs little from that of the dominant contribution...in the absence of a dominant component, combine them by taking the square root of the sum of the squares. This gives the combined standard uncertainty” (UKAS, 2007, [The Expression of Uncertainty and Confidence in Measurement](#)).

If there are more than two major contributors to variability, do the sum of squares with all variance (SD squared) terms included to see if including the relatively low magnitude SDs changes the results appreciably.

Beyond the issue of having more than two major contributors to variability, there are other issues that might make the 1/5th rule not work optimally in every situation. For example, in some types of reconnaissance-level sampling, it is impossible to collect enough samples to accurately define natural variability (either temporal or spatial) within the time frame and funding available. On the measurement scale of concern, unless one is looking over a long time period and multiple [QC](#) samples, one often does not have a large enough sample size to accurately estimate a standard deviation for measurement [precision](#). In fact, at first one typically just has sample size of two (and thus difference is expressed as a relative percent difference---[RPD](#)). Why is this important? When either the numerator or the denominator (and especially both) are not good estimates of the SD, one can't accurately judge the 1/5th threshold. The 1/5th rule depends on good estimations of the standard deviations. Standard deviations are typically not well estimated at small sample sizes (below 25-30 and especially below 7-10) or when the values used to generate the SD for true environmental heterogeneity do not represent the full range of conditions of the representative [Target Population](#) being sampled.

If sample sizes are too low or if calculated SDs are not representative of the target population, these faults should be corrected before putting too much weight on the results of the 1/5th rule of thumb.

This factor of 5 (expressed as a SD) is generally consistent with other signal to noise rules of thumb. Most of these rules of thumb state that (for accurate measurement) a signal should typically be 3 to 10 times greater in magnitude than noise (see [Part B](#) for more detail and several examples). Statements such as "errors in the analytical measurements should be no greater than the natural variability of the parameters of interest" should be rejected since such errors should usually be at least 5 times lower (when expressed as SDs).

What about other summary statistics used for variability? Can one also see if either a coefficient of variation (CV, the standard deviation divided by the mean) or a relative standard deviation (RSD = CV*100) is 5 times lower than their counterparts, when estimating true environmental variability (on measurements of different samples)? No, these are different. The 1/5th rule of thumb should be used with SDs only. Using this rule for other summary statistics produces different results.

Calculate Monitoring Design Sensitivity

Why determine detectable differences in summary statistics based on measurements of different samples? Generic VS guidance has suggested that networks “List the specific, measurable objectives for each vital sign selected for monitoring, and wherever possible, give the [threshold](#) value or “trigger point” at which some action will

be taken” (NPS. 2004. [Outline for Vital Signs Monitoring Plans](#)). To be precautionary in preventing major impacts or even a resource collapse, monitoring networks typically have to be able to detect a change smaller than the chosen thresholds or trigger values (see more detail below).

Effect Sizes (ESs) Based On Multiples of the Standard Deviation

The effect size (ES) in social sciences is usually expressed as a % of the true standard deviation rather than original units of measure. It is usually the minimum detectable difference between means (or between a mean and a water quality standard or other benchmark) divided by best estimate of the true (but unknown) population standard deviation. That best estimate is most often a pooled standard deviation that covers a broad range of conditions. The pooled standard deviation (SD) for two or more samples is essentially the square root of the average variance (Widener University Descriptive Statistics home page includes a [full equation for a pooled SD](#)).

The ES result is the magnitude of change expressed as the number of standard deviations. One then multiplies the result x 100 to get the result expressed as a percent of the standard deviation. So the final ES is in % units rather than original units of measure. One ES advantage is that comparisons of effect sizes for different measures or vital signs can easily be made between different vital sign measures and time frames, since all ESs are expressed in SD units. Also, no initial estimates of variability are needed; a big advantage if one is making initial calculations before any credible estimates of variability are available.

In psychology, a “large” standardized effect size (ES) would be considered to be an 80% change in Standard Deviation (SD) units (Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*. Lawrence-Erlbaum, Hillsdale, N.J.). This would often be a small change in field biology or field water quality scenarios.

Also, environmental variables are often not normally distributed and sample sizes are often small, so ES calculations are usually used only for very rough initial estimates, for comparing effect sizes **between indicators** and for looking at effect magnitudes from different angles (not using original units).

A major drawback of using ESs in SD units is that if one chooses a certain ES magnitude (say 80% for example), one will choose the same sample size regardless of the accuracy or reliability of the measuring instrument or the true variability of what is being measured. This is clearly not ideal, and one more reason to also look a power or detectability in original units as soon as possible. One should use power prospectively, put science before statistics, and do pilot studies [[Lenth, R. V. 2006. Java Applets for Power and Sample Size](#) (Advice Section)].

Although some investigators also use the phrase effect size when talking about original units of measure, to avoid confusion, herein we use it only when referring to differences as multiples of the standard deviation, typically to used only very early in the monitoring planning process or when trying to standardize differences between indicators.

Minimum Detectable Differences (MDDs) in Original Units

A minimum detectable difference (MDD) in original units of measure is typically a minimum detectable difference between means (or medians, or other summary statistic) in the current sample compared to either 1) a summary statistic from a different sample (collected somewhere else in time or space), or 2) a water quality standard or other benchmark.

A good example of a NPS Vital Signs network calculating and factoring in minimum detectable differences is the [NCPN Freshwater Protocol](#). Other water quality protocols narrative drafts (San Francisco, Pacific Islands, and Great Lakes Networks) are also at various stages of documenting minimum detectable differences.

Choosing on the monitoring design scale of concern is usually done in an iterative manner. If one does not already have an established and reasoned goal, first choose an initial MDD that seems reasonable. For example a network might decide to pick a 30-40% change (in original units of measure) over a one year period as an initial MDD change in a mean as a starting point. Next, have the network quantitative ecologist run the numbers ([step 12](#)) to see if the monitoring design will be able to detect a change that small. Often initial decisions on sample sizes, sample numbers, sample locations, and detectable differences simply will not work and adjustments in one or more of these factors (and/or in alpha or beta) are needed to improve detection probabilities.

There are a few isolated cases where the NPS has logical reasons for the need to be able to detect very small change, such as a 5% MDD as change in means in original units (not SDs). For example, for air quality goals, a 5% change in visibility was shown in human studies to be "perceptible." The Clean Air Act states that "visibility impairment" is defined as "any humanly perceptible change in visibility from that which would have existed under natural conditions." (USFS, FWS, and NPS. 2000. [Federal Land Managers Air Quality Related Values Work Group Phase 1 Report](#)).

However, in open aquatic habitats, it is usually it is difficult to detect changes that small in means or original units, so it is more common to try to detect 20-40% differences, sometimes from year to year or sometimes over a stated number of years.

In deciding how big of a change or difference monitoring needs to be able to detect, it is helpful to consider the type of park(s) being monitored and park-specific management goals. In the variable worlds of aquatic biology and water quality, detecting a 5% change, either as a MDD or ES, would usually be impossible or require so many samples that it would be prohibitively expensive. Keep in mind that the smaller the MDD or ES one is trying to detect, usually the more samples one would have to take (more costly) and the more difficult it is to find a strata where the variability is low enough to allow detecting such a small change. Therefore, it is usually not advisable or often not possible to detect extremely small changes (1-5%). Consider the following more typical scenarios:

Scenario 1: The resource to be protected is an endangered species or a very highly prized and rare resource in a relatively pristine area of a park having natural resource protection as a key goal: In this case, a park resource stewardship plan might logically call for a relatively high degree of protection and management precaution. Such a park might even choose to protect such a resource very

stringently. Anything above a negligible or even a detectable impact of concern might not be considered acceptable relevant to desired conditions (DCs). The minimum detectable difference (MDD) the park might want to detect might be a relatively stringent 5-25% change in means in original units of measure. In SD units, the ES size the NPS might want to detect might be as small as a 10-30% change when the change-magnitude units are the number of standard deviations expressed as a %.

Scenario 2: The resource to be protected is a population of an aquatic species at a typical national park, but the species is very common in the region and/or nation (for example, a bluegill sunfish). In this case, the park might designate a less-stringent MDD, such as a more common 20-50% change in means, or an ES change-magnitude of 30-70%, when the units are a percentage of the magnitude of the standard deviation.

Scenario 3: The natural resource to be protected is general water quality or a population of a common aquatic species at a historical park in a highly urbanized or highly farmed area. Here the water quality and aquatic habitat is such that there is little or any hope of ever reaching “unaffected by modern civilization” status, and this is reflected by state water quality standards and biocriteria that are less stringent than one would find in less-impacted areas of the country. Perhaps the biota to be protected is short-lived and highly variable even in pristine areas. In this scenario, the park might decide that only larger changes can or need be detected. The park might therefore want to be able to detect a MDD of a 40-80% change in means, or an ES change of 70%-90% when the units are a percentage of the magnitude of the standard deviation.

The examples above are mentioned only to give monitoring planners a very rough idea of some typical ballpark (starting-point) ranges of values. Whenever one has first developed a logical and defensible park-specific MDD or ES, of course those values should be documented and used instead of the examples above.

The NOAA [National Estuarine Eutrophication Assessment](#) (NEEA) monitoring program specified the following starting-place change magnitudes for [Submerged Aquatic Vegetation \(SAV\)](#):

A change in spatial coverage of the SAV beds was considered very low from a 0-10% change, low from a 0-25% change, moderate from a 25 - 50% change, and high for a >50% change. However, NOAA is now thinking of alternatives, since some believe a 25% change might in fact be very significant (Suzanne Bricker, NOAA, Personal Communication, 2007).

What if the park simply has no idea whatsoever how big of a MDD they need to be able to detect to protect important resources? Try detecting changes or differences of the magnitudes mentioned above as a starting point.

For some indicators, the ability to detect a 20% change in means in one year or even multiple years might require too many samples and exceed the budget. A network

could consider the adaptive-management approach that Channel Islands National Park adopted as a starting point for VS monitoring. That park adopted a preliminary goal of being able to detect a 40% change in means from year to year, with alpha (Type I error, the polluter's risk) of 0.05 and beta 0.2 (power 80%, the Type II error, the conservationist's risk), with the stated idea that the values could be changed later if needed, as lessons were learned. An exact quote, which provides thought process documentation, from the CHIS document was

“People who use the monitoring information made these guidelines explicit, based largely on concerns about cost and accountability for the nation's heritage. They determined that the park could not afford to detect 5-10% changes and could not afford not to detect 50% losses of critical resources, such as endemic species. This 40% goal was a pragmatic compromise between cost and risk. It was an attempt, in an adaptive management scheme, to balance scientific credibility and practicality that could be tested and modified in response to experience.” (Davis, G. E. 2005. National park stewardship and ‘vital signs’ monitoring: [A case study from Channel Islands National Park, California](#). Aquatic Conservation Marine Freshwater Ecosystems 15:71-89).

In common situations where one is not trying to protect an endangered species or something else especially rare or valuable, it is uncommon to try to detect a biological effect size smaller than a 20% change in means, especially in only one year in the highly variable world of water column variables. The 20% default is often adopted as a “de minimis” (the law cares not for little things) starting point when one cannot logically come up with something better, though in water quality it might more often be over several years rather than one.. A 1992 paper suggested that it was difficult to find cases where a state or federal regulatory agency had prosecuted anyone for a biological effect size of less than 20% of the mean on non-human or non-endangered species. This was true regardless of whether the population, community, or ecosystem level was being considered (Suter, G.W. II, A. Redfearn, R.K. White and R.A. Shaw. 1992. Approach and strategy for performing ecological risk assessments for the Department of Energy Oak Ridge Field Office Environmental Restoration Program. Martin Marietta Environmental Restoration Program Publication ES/ER/TM-33, Environmental Restoration Division Document Management Center Environmental Report (ER), Environmental Sciences Division (ESD) Publication 3906, Oak Ridge National Laboratory, Oak Ridge, TN, pp. 8-9).

Funding limitations should not be an excuse to monitor a large number of sites (or a large number of measures) poorly. It would be better to monitor fewer sites and fewer vital signs well). Or, as said in more technical terms in the 2006 [Southwest Alaska Network phase III Monitoring Plan](#), “It is better to gather sufficient data on a smaller area of inference than inadequate data on a larger scale of inference” and “It is better to gather data of sufficient quality on fewer vital signs than insufficient data on many of them”

Again, in water quality, one can sometimes reduce variability by carefully picking integrator variables (for example, benthic macroinvertebrates) and by narrowly defining [Target Populations](#) (for example the populations in riffles only, or on snags only), in

narrow index time-periods only). This tends to reduce variability compared to randomly sampling water column variables in all habitats at all times of year.

Reducing variability with narrow definitions of strata, target populations, and extent of inference, can help one achieve more uniform magnitude of variability between-years and thus help one pass the straight-face (common sense) test when claiming that the results from samples from multiple years can be lumped before estimating a proportion, and still represent a valid single-sample. In done right in context, this can facilitate the justification of rotating panel designs that can reduce the number of samples needed per year. Reducing variability also helps one detect changes of a size of concern.

Another strategy is to move a continuous monitoring sonde around on a rotating panel basis to get required sample sizes and to document temporal variability. In concert with oft-repeated generic NPS vital signs guidance, networks should design monitoring that produce credible information with available funding, but identify areas for expansion. As we deliver value to parks and increase our partnerships, additional resources may be available to support a larger program.

No matter how the quantitative levels are developed, once the MDD or minimum detectable ES is developed as our best quantitative estimate of the change that the park or network would like to be able to detect, a reality check comparison should be made between that value, the current condition, and individual water quality parameter or indicator targets needed to help achieve DCs. What is the amount of change it would take to get to different levels of compliance with water quality standards or to get to levels that would cause a resource collapse (if known)? All such goals should logically reflect park management goals. In the park planning process, such a comparison might be made in the newer-style NPS resource stewardship plans.

Monitoring networks should usually try to be precautionary by using starting beta levels of 0.05 to 0.1, if standard null hypothesis significance testing (NHST) is envisioned. The other alternative if sample size is smaller than 30 would be to estimate needed sample sizes and power using the more precautionary [inequivalence](#) test rather than standard NHST.

In any case, the starting point MDD can be changed to a different % based on park protection goals or ecological and other lessons in an iterative (adaptive management) fashion. For example after initial pilot data is gathered and initial MDDs are estimated, these may change later as more data and real-world experience is accumulated, and/or as statisticians later provide more advanced ways to look at MDDs, power, and needed sample sizes.

9) What Initial Statistics Will Be Used?

Do the questions to be answered call for detecting A) a difference between two means, B) a difference between a mean and a water quality standard, C) a trend (over how many years), or D) an estimation such as a confidence interval (CI) about a mean or a proportion? Will parametric or nonparametric statistics be used? Are there [nondetects](#)? The answer to these questions will help drive which statistics will be used and how sample sizes are calculated. See [Which Test Do I Use?](#) Statistics chosen should be documented in a [data analysis SOP](#).

10) Choose Desired Confidence/Detection Probability (Power = 1-beta)

What is the degree of confidence we want to have in detecting our chosen change magnitude of concern? This is a decision for park and other NPS staff to make with the help of the network quantitative ecologist. Do we want to have 80%, 90%, or 95% confidence that we can detect a defined change-magnitude of concern? The decision may vary with rareness and special value of the resource being protected, how pristine the area is, or the variability of the vital sign or measure. For a measure that is highly variable even in a variability-reducing stratum, it may difficult or impossible to detect a 20% change over one year or even multiple years with 90% confidence.

Typically the network would work with a NPS quantitative ecologist (and later with a statistician) in an iterative manner to decide workable probability of detection percentages. In other words, once an initial decision has been made about desired probability of detecting the effect size of concern, it is typically necessary to run the numbers to determine the actual probability of detecting an effect size of concern given the monitoring design, the variability of the parameters, the statistics chosen, and other factors discussed herein. If the detection probabilities turn out to be unacceptably low, changes in the overall monitoring design, analytes to be measured, sample sizes, sample locations and strata, sample timing, etc. may have to be made. It is common to have to repeat sample size calculations several times before satisfactory compromises between power, budget, and choice of effect sizes can be made.

Use past data to get the best estimate you can find for standard deviations or variances to use in the power and sample size calculators. The dreaded post-hoc power prohibitions are not an issue here because the scenario is not the situation where one is “after the fact, doing a power analysis for the *effect observed* in the study.” This scenario of planning future monitoring is more acceptable because. (Ken Gerow, University of Wyoming Statistics Department, Personal Communication 2007):

- (1) Any data set can be thought of a “preliminary” in that it stands as fodder for future work.
- (2) It is indeed better to have (our best available) estimates of SD and so on (which requires preliminary data).
- (3) Post-hoc simply means (“after”) which sounds in fact like the *right* thing to do (if you take my points (1) and (2) as valid. So it is perfectly fine (good, even) to do post-hoc power analyses in this particular scenario of looking at past monitoring estimates to get the best possible estimates of variability to help plan for future monitoring.

In standard null hypothesis significance testing (NHST) scenarios, the degree of confidence we have in detecting a change of a certain size (a defined minimum detectable difference--[MDD](#)) is the same as statistical power, and correlates very highly with sample size. That is because at higher and higher sample sizes, lower and lower differences can be detected.

In a standard NHST scenario, a ninety % confidence might also be expressed as a power of 0.90 (1-beta when beta is 0.10 or 10%). Expressed as a %, we would say the statistical power is 90%. Power is most useful in pre-monitoring planning, to be able to state (for example) that the study was designed to be able to **regularly** detect a certain

magnitude of change (say a 20% change in means) a certain % (say 90%) of the time. Another way to express it is that we have designed the study before doing it to limit beta to 10%, so 90% of the time we will not make a false negative conclusion that there is no effect of the stated size when in fact there is an effect of that size.

The McBride probability of detection calculator clarifies that power (1-beta) is treated the opposite for standard NHSTs vs. the more precautionary (especially at smaller samples sizes, see second paragraph below) [inequivalence](#) tests. In inequivalence testing, the burden of proof is shifted towards the polluter and away from the conservationist, though at very large sample sizes with good control of power, standard NHST (tests) can become quite precautionary ([McBride calculators](#)):

In most cases the “detection probability” is akin to the test's "Statistical Power" (the probability that the tested hypothesis of no difference will be properly rejected when it is in fact false). For testing the hypothesis of inequivalence, the analogous concept is the test's "Operating Characteristic" [the probability that the tested hypothesis (inequivalence) will not be falsely rejected]. So there are two differences: 1) inequivalence vs. equivalence, and 2) whether or not the tested hypothesis will or will not be rejected. These mental gymnastics, however, allow the inequivalence “Operating Characteristic Curves” to be fully analogous to the power curves for the more familiar null hypothesis significance testing (NHST). As pointed out by McBride in a help file off his [calculator](#), the operating characteristic can be thought of as “the frequency with which a correct hypothesis (of inequivalence) will be accepted”, and to accept the inequivalence hypothesis is to detect a truly important difference, which requires high values for the “operating characteristic.”

At sample sizes below 25-30, an [inequivalence](#) test should be the precautionary default-first choice of the National Park Service, which after all, is in the business of protecting rare and special resources for future generations. When sample size is above 30, the more familiar standard NHST tends to have a power no less than 90-95%, and if properly executed and interpreted, therefore becomes an acceptable alternative as one line of evidence in multiple lines of evidence assessments. More detail:

As clarified in Section 5.3.3 of the [McBride statistics book](#), at small sample sizes (such as sample size of less than 25) and an equivalence interval effect-size of the magnitude of half of the standard deviation, the standard NHST (test) is neither as permissive as the test of the equivalence hypothesis nor as conservative as a test of the [inequivalence](#) hypothesis. However, at large sample sizes (such as $n = 50$ to 100, the null hypothesis test (NHST) routinely becomes very conservative, routinely detecting even small differences, and its detection probability is everywhere (different combinations of effect sizes and sample sizes) higher than that of the inequivalence test. If these concepts seem too difficult, have your local applied water quality statistician help you sort them out.

If hypothesis testing is not done, but estimation (say of a confidence interval about a mean or proportion, see [step 12](#)) is done instead, the degree of confidence is typically expressed by the magnitude of the confidence interval rather than beta, but care should still be taken to assure adequate sample sizes (See “[Confidence Intervals about Means.](#)”

11) Choose Significance Level (*alpha*)

What is the desired probability (1-significance level = $1-\alpha$) to avoid falsely detecting a change or difference of the magnitude of the [MDD](#) (probability of type I error, the polluter’s risk)?

Too often what looks like science---choosing an error rate for a statistical test for water quality assessments---is an unrecognized public policy decision (Shabman, L. and E. Smith. 2003, Implications of Applying Statistically Based Procedures for Water Quality Assessment.” *Journal of Water Resources Planning and Management*, 29(4): Pp. 330-336, see rest of quote and related discussion on page 176 of [McBride statistics book](#)). In the p-value culture of many journals in the past, significance level (α) has traditionally been set at 0.05. In the p-value culture of many journals in the past, significance level (α) has traditionally been set at 0.05. Although recently there have been many articles explaining why automatically doing this (and over-reliance on p values and NHST tests in general) is not a good idea, the culture persists. For more information on the problems, see [Part B](#) and some of the many recent discussions, such as

Stefano et al. 2005, [Effect size estimates and confidence intervals](#)

S. Rigby. 1999. [Getting past the statistical referee: moving away from P-values and towards interval estimation](#)

B.D. Mapstone. 1995. Scalable Decision Rules for Environmental Impact Studies: Effect Size, Type I, and Type II Errors. *Ecological Applications* 5 (2).

J. Gliner et al. 2001 [Null Hypothesis Significance Testing: Effect Size Matters](#)

When doing [inequivalence](#) testing, $\alpha = 0.05$ is a fine default. For other types of testing, the NPS is typically even more concerned with false negatives (wrongly concluding no impact or no change when a change or impact has, in fact, happened) than false positives (wrongly concluding impact or change when a change or impact has not happened). Therefore, network quantitative ecologists may at times specify small beta levels, even if they are smaller than α (see [Part B](#) for more details).

Keep in mind that equivalence testing is not all that precautionary (the conservationist risk, type II errors) unless sample size is high ([Practical Stats January 2007 Newsletter](#)). Accordingly in the NPS, we wouldn’t ordinarily recommend equivalence testing, but would instead tend to favor inequivalence testing, which tends to be more precautionary at small sample sizes than standard NHST testing [especially in null hypothesis scenarios where only α (and not beta)] are constrained to very low levels (such as the traditional 0.05 for α).

Unless one was using an inequivalence test, instead of blindly insisting that alpha always be 0.05 according to tradition, monitoring planners might choose 0.1 or 0.2 for alpha. Beta might then be 0.05 or even 0.01. As a resource conservation agency which typically desires to be precautionary in protecting rare and/or highly valued resources, the NPS should not be less worried about beta (the conservationist's risk) than the alpha (the polluter's risk in standard null hypothesis test). Therefore, consider choosing [inequivalence](#) tests rather than standard null hypothesis tests at smaller (less than 25) sample sizes

12) Use Simple Calculators to Make Initial Estimates of Required Sample Sizes

Due to recent breakthroughs, it is now easier for quantitative ecologists to use sample size equations and calculators on the Internet to get a rough idea of sample sizes needed. Power should not be ignored, so be sure that any minimum detectable difference ([MDD](#)) calculator used has inputs not only for alpha, but also for beta (1-power) and for the standard deviation (SD) or variance). If ES calculators are used as a first step, be sure the calculator has inputs for alpha AND for beta (1-power) and follow up with MDD calculators to look from a better angle when a SD is available.

One typically does sample size calculations in an iterative manner, trying various samples sizes until the answer stabilizes (rounds up the same whole number) and/or by playing what-if games. If one already has a starting sample size, the McBride [probability of detection calculator](#) can also be used to estimate the effect size that can be detected with various sample sizes, probabilities, stated significance levels, and various types of tests. Start by clicking on the effect size button and then choose either one or two groups.

CAUTION: There are many sample size and power calculators on the Internet. Before using them, we suggest checking them against Zar's examples to make sure you can get the right answers. Some of the Internet calculators appear to use the wrong equations and/or give the wrong answers. One used the [two-sided](#) critical t-value instead of the (correct) [one-sided](#) critical t-value for power. Also, it is easy to misunderstand some of the input variables or their format, which is why detailed step-by-steps are given in [Part B](#). As will be repeated for emphasis, hypothesis test sample size estimators that don't take into account power (1-beta, or analogous probability of detection for [inequivalence](#) tests) rates should not be used unless otherwise justified.

Step-by-step examples on how to use some of the more user friendly Internet calculators to get the same answers as the Zar examples are summarized below and in more detail in section IV-C of [Part B](#).

Simple sample size calculators provide just rough (but far better than nothing) starting point estimates of needed sample sizes. Some would argue that the same is true for more complex simulation approaches.

Perform Different Initial Simple Calculations Depending on the Scenario:

We recommend estimating sample sizes needed with a multi-step approach. No matter how one is estimating required sample sizes, it is important to keep documenting and checking assumptions at each step along the way. Calculations for sample size requirements, like all inferential techniques, are based on certain assumptions. Be sure to discuss assumptions in eventual discussions with an applied statistician ([step 16](#)).

The first few steps listed below relate to hypothesis testing, since these tests are still commonly used and since many are familiar with them and with the Zar equations (details below). This does not imply that standard null hypothesis significance testing (NHST) should be a first choice for analysis. In fact, the standard null hypothesis test (especially when power is not controlled) is most often not the only good choice or the optimal choice for ecological field work.

For a recent balanced review of the pros and cons of NHST versus common alternatives, see R. S. Nickerson. 2000. [Null hypothesis significance testing: A review of an old and continuing controversy](#). *Psychological Methods*, 5:241-301. Among the conclusions therein: 1) All statistical tools have use in our tool box, but all are also subject to misinterpretation and misuse, 2) There are many choices beyond just Bayesian or not, that two-way split is much too simple, 3) confidence intervals (and most other alternatives, including Bayesian alternatives, are also prone to misuse and misinterpretation, 4) in spite of vehement criticisms of NHST (many of which were really about misuse rather than about proper use), it remains the most popular tool used in psychology papers, 5) There is a need to be careful with terminology and what is said in interpretation. For example, many of what is said about alpha, beta, and confidence intervals, even in some statistics textbooks, is either at worst flat-out wrong or at best potentially misleading, 6) one reason for doing pre-hypothesis test calculations (including power and needed sample sizes) is to assure adequate sample sizes [which usually proves helpful when data is used for other purposes (confidence intervals, etc.) too], and 7) Although the likelihood that a true null hypothesis will be rejected does not increase with the sizes of the samples compared, the likelihood that a real difference of a given magnitude will result in rejection of the null hypothesis at a given level of confidence does. It is also the case that the smaller a real difference is, the larger the samples are likely to have to be to provide a basis for rejecting the null. In other words, whether or not one assumes that the null hypothesis is always or almost always false, when it is false the probability that a statistical significance test will lead to rejection increases with sample size.

Although aimed at psychologists rather than ecologists, among the many other quotes Nickerson (op.cit. just above) made that were of particular interest were several from Abelson, who (through a series of papers) acknowledged issues with null hypothesis significance testing (NHST) but also defended certain judicious and careful use of NHST under certain well justified scenarios (proper calculations, checked assumptions, only one line of evidence, etc.). Among these various Abelson quotes were the following (see Nickerson op.cit. for full citations):

- 1) If a legal case were being brought against the significance test, the charge here would be that the test is an 'attractive nuisance,' like a neighbor's pond in which children drown. It tempts you into making inappropriate statements,

- 2) All statistics, in his (Abelson's) view, should be treated as aids to principled argument. He saw NHST as not the only, but an essential, tool in the researcher's kit: "Significance tests fill an important need in answering some key questions, and if they did not exist they would have to be invented."
- 3) Even though a single study cannot strictly prove anything, it can challenge, provoke, irritate, or inspire further research to generalize, elaborate, clarify, or to debunk the claims of the single study.
- 4) Whatever else is done about null-hypothesis tests, let us stop viewing statistical analysis as a sanctification process. We are awash in a sea of uncertainty, caused by a flood tide of sampling and measurement errors, and there are no objective procedures that avoid human judgment and guarantee correct interpretations of results.

The Nickerson summary (op.cit.) also contains many other quotes of others plus some original statements that are noteworthy. Among the interesting statements of special interest to those doing biomonitoring were the following (see Nickerson, op.cit for original citations and context):

1. Estes has pointed out that statistical results are meaningful only to the extent that both author and reader understand the basis of their computation, which often can be done in more ways than one; mutual understanding can be impeded if either author or reader is unaware of how a program has computed a statistic of a given name.
2. Some point out that many of the criticisms of NHST are not so much criticisms of NHST *per se* but criticisms of some of its users and misuses. It is not the fault of the process, they contend, if some of its users misunderstand it, expect more from it than it promises to deliver, or apply it inappropriately.
3. Beta is simply the probability of failing to reject the null hypothesis, given that it is false.
4. Statistical tests, at least parametric statistical tests, invariably rest on certain assumptions. Student's t test and the analysis of variance (ANOVA) involve assumptions regarding how the variables of interest are distributed. Student's t, for example, tests the hypothesis that two samples of measures of interest were randomly drawn from the same normally distributed population with a specific variance. However, neither the shape of the population of interest nor its variance is typically known. It is often simply assumed that the distribution is normal and the variance is most often simply estimated from the sample values. For application of the test to be legitimate, the samples should be normally distributed and they should have roughly equal variances.

5. Procedures have been developed for determining whether samples meet the requirements and for transforming the raw data in certain ways when they do not meet them so they will. The degree to which researchers ensure that data satisfy all the assumptions underlying the significance tests they apply to them is not known; my guess is that it is not high.
6. The width of a confidence interval is generally a random variable, subject to sampling fluctuations of its own, and may be too unreliable at small sample sizes to be useful for some purposes.
7. The use of point estimates and confidence intervals ("error bands") predates the development of NHST. Cohen's (1994) surmise is that the reason they are not reported is that, at least when set at 95% to correspond to a .05 *alpha* level, they typically are embarrassingly large.

Before doing statistical tests, looking at data from different angles, as part of exploratory data analysis (EDA), including summary statistics and simple plots, is always a good idea as it sometimes reveals important information not otherwise immediately apparent, as well as suggesting which types of tests or other analyses might be optimal.

Looking at results from subsequent additional tests and analyses is a further way to look at the results from different angles. Try to assemble numerous lines of evidence. For example, when looking at trends in pH in water, one might perform a seasonal Kendall nonparametric test for trends, and plot the data to see if there is a hint of a slope (and/or do a regression test to see if the slope is different than zero), and do a t-test (or nonparametric analog for parameters other than pH which is already transformed) to compare say one 25 year period to the next 25 year period, and do a sample size estimates to make sure sample sizes were adequate to ensure needed amounts of statistical power (detection probabilities). If sample sizes were high enough, one might also perform [inequivalence](#) tests. The point is that if several or a majority of different lines of evidence support a trend for pH, these multiple lines of supporting evidence would be more convincing than any one line of evidence by itself. Null hypothesis tests have come under fire in recent years, but if done with sufficient power as one of several lines of evidence (rather than being considered definitive as stand-alones), they can add information value. Two lines of supporting evidence would typically be considered more convincing than one, three more convincing than two, etc.

Even when inequivalence tests or other good options involving controlling power are chosen, single test results are often not definitive by themselves and typically should be used only as one of many lines of evidence considered in ecological resource management decisions.

Nevertheless, parks often want to compare results from one site to another or from one time period (say one 5 year period of drought) to another time period (for example 5 relatively wet years). Sometimes parks even want to compare one year to the very next. More often, if they visit plots two years in a row and then rest them for several years in a rotating sampling approach, they might wish to compare the average of these two year periods to the next two year sampling period. In several of these scenarios, the first few items below are relevant for estimating needed sample sizes.

Observed to expected (O/E) ratios tend to be normally distributed and sample sizes can be calculated with standard calculators such as the ones summarized below. If a NPS VS monitoring network is developing a network assessment or network O/E predictive model, the number of sites is the level of replication and emphasis should be placed on sampling a sufficient number of sites to improve regional. Replication at a site or reach may be required for the following reasons: 1) to find variability at a single site, 2) to get confidence interval size down to a reasonable % of the O/E ratio, and/or 3) to improve the ability to detect small changes.

Note: Other things being equal, pristine sites, or those with little human influence, tend have smaller variability than impacted ones, which helps when one is trying to detect small changes (for more information, see S.J. Nichols, W.A. Robinson R.H. Norris. 2006. Sample Variability Influences on the Precision of Predictive Bioassessment. Hydrobiologia).

Before using calculators, it is very important make sure you understand exactly how to input variables need to be formatted and entered and check some example problems where the correct answer is known before proceeding.

It is also a good idea to perform multiple calculations from the options listed below to “look at the issue from multiple angles” and to see if they are all close. If they are not, an input variable may be wrongly formatted. If they are close, take the highest sample size estimate to be precautionary, in the absence of a better rationale.

Sample Size Calculations for Nonparametric Procedures

Most of the sample size (and related power) calculators use the t-distribution and assume normality (of the sampling distribution of the means, *not* of the distribution of values themselves). What if assumptions are not met and nonparametric tests will be used? Nonparametric sample size and power estimators are not easily available.

Most experts have suggested that the power and sample size requirements for nonparametric tests are usually not all that different than those for parametric tests. Some suggest that one could use the parametric calculators for first estimates of sample size and then add a small amount (perhaps 5-10%) to be precautionary if one is unsure of the normality of the distribution being sampled. The initial sample size estimates are usually rough anyway, and if one wanted really exact values, and if one had a large sample size truly representative of the [Target Population](#), one might go to more complex methods (such as simulation) to estimate sample sizes.

Others see adding a small amount to the sample sizes for use in nonparametric tests as unnecessary, since field environmental data is usually badly skewed. They point out that when the assumptions of the t-test are badly violated, the Mann-Whitney test has **more** power than those tests that require assumptions of normality (Zar, op.cit). The Wilcoxon Sign Ranks test for two independent samples has about a 95% efficiency compared to a t-test when the distribution is normal. See [Part B](#) for more additional documentation and detailed information for different multipliers for different nonparametric tests in the unlikely case that one runs across normally distributed distributions in environmental sampling

However, as pointed out by Johnson in 1999, if available past data is marked by small sample size or questionable [representativeness](#) (almost always the case), even for normally distributed data, the parametric sample size calculators, even those with inputs for alpha, beta, and delta (detectable differences) often underestimate required sample size, partly since available estimates for variance are often not a good reflection of true variance in the underlying [Target Population](#) (Johnson, D. H. 1999. The insignificance of statistical significance testing. *J Wildl Mgmt* 63:763-772.). Therefore, increasing the sample size a bit to be precautionary is not a bad idea even for parametric sample size calculators.

A related strategy that EPA uses in CERCLA sites to prevent underestimating variability is to use a 80 or 90 percent upper confidence limit for the estimate of the standard deviation rather than an unbiased estimate to avoid underestimating the true variability (EPA 2002, [Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites](#))

Many investigators use the t-distribution calculators for data that is log-normally distributed and then log transform environmental values before using parametric tests, especially when sample size is above 20-30. However, this should not be an automatic choice. Sometimes nonparametric analyses are a better choice, and simulation can sometimes do a better job of estimating needed sample sizes. Also, be careful to avoid back transformation bias. If the mean is the main focus and the test is to be done on transformed values, then run the sample size estimators with transformed values, and don't report back transformed means, standard deviations, minimum detectable differences [MDDs](#) in means, or variances. MDDs in geometric means (medians) can be used.

Note: Again, the key point is that back transforming a log mean gives the geometric mean, which estimates the median (not the mean!) of the original units assuming the logs are symmetric. If logs are nearly symmetric, sample size calculations in log units will be a good approximation to true nonparametric sample size estimators (Dennis Helsel, USGS, Personal Communication, 2006).

The Helsel and Hirsch text book (Helsel, D.R. and R.M. Hirsch. 2002. [Statistical Methods in Water Resources](#). US Geological Survey Techniques of Water Resources Investigations) does not go into sample size calculations in detail. However, chapter 4 provides good detail on using nonparametric hypothesis tests, even with small sample sizes.

Before a Good Estimate of the Standard Deviation is Obtained:

If a hypothesis test is to be used, perhaps as one line of evidence in a multi-lines of evidence approach, we suggest starting sample size estimation with the [McBride detection calculator](#). This is a good first step since: 1) it allows one to make initial estimates of sample sizes if there is no (or no good) estimate of variability, and 2) it allows one to look at probabilities of detection of various effect sizes in not only NHST t-tests but also in the generally precautionary and more optimal [inequivalence](#) tests, and the less precautionary equivalence tests. It also allows for either paired or not-paired

sampling estimates. Impacts may change not only means but also SDs (more impact often leads to more variation). Therefore, it is not a bad idea to pay close attention to changes in standard deviations (SDs) and in changes expressed as a percent of a SD. In the McBride calculator, remember to express effect size as a percent of the standard deviation in the calculator input. So, if $ES = \text{difference}/SD = 0.69$, express the effect size as 69%. Be sure the SD used is a best estimate of true heterogeneity of different samples, not generated from repeat sampling of a single sample. For a plain-language step-by-step for how to use the McBride calculators, see section on [Inequivalence](#) testing.

After a Good Estimate of the Standard Deviation is Obtained:

Once one has a decent estimate of a SD, do another quick calculation based on Zar's equations (Zar, J. H. 1999. Biostatistical Analysis. Prentice Hall, Upper Saddle River, New Jersey, USA). Many of these are discussed below.

Sample Size Needed to Detect a Defined Difference between Two Means

The following section explains how to compute the sample size needed to detect a pre-determined magnitude of minimum detectable difference (MDD, expressed in original units of measure) between two means. Here one assumes that someone has made a decision that they want to be able to detect a difference of a given magnitude (say 20 or 40 or 50%) between two means. The example here assumes the two samples being compared have similar amounts of variability (= similar variances or similar standard deviations) and equal or very close to equal sample sizes.

For this scenario, one can use Zar's (1999) equation 8.22. Planners may be able to use an Internet calculator after confirming they give the same answer as Zar provided

Using the Gerow Calculator to Get Zar's Answer:

When checked in 2006, the [Gerow calculator](#) gives the same answer (45 for each sample) as Zar gives in example 8.4 (page 134, Zar 1999) for equation 8.22. Here is the step-by-step method for the Gerow Calculator to replicate the Zar answer of 45 (starting at the upper left hand corner of the Gerow Calculator):

1. Set alpha to 0.05 with the slider bar, then immediately below that
2. Toggle to the [two-sided](#) test option, then move down and
3. Toggle to the "increase" option, then move down and
4. In the null mean box, type in one mean (use 1 for Zar example)
5. Then move the "Size of Difference" slider bar until 1.5 appears in the box for alternate mean (this uses the Zar example), then just below
6. Move the arrows in the "Variation Choices" Box until the Constant SD choice appears (for equal variances), then just below that
7. Type in 0.721 in the "estimate of SD" box,
8. Below the box where you have entered 0.721 there is another box, type in zero, then move to right and

9. Move the arrows in the spin button for “Design Choices” until “independent, equal sample sizes” (the Zar choice) appears, then just below that
10. Move the Sample Size Slider with the arrows until the desired power (90%) appears in the power answer just above (and a bit to the right) of the center of the Design Choices Box. In this case either sample sizes of 44 or 45 correspond to 90% power, so choose 45 to be conservative (45 is also Zar’s answer, so one can see that input choices above were correct). Thus the highest sample size at which power is still 90% and not 91% is the answer to be used (keep toggling until 91 appears, then toggle back one step until the highest number associated with 90% power, in this case 45, appears).

Note: The vague notion that estimates of standard deviations and means both tend to become more accurate as sample size increases **is not** a reason to choose “Standard Deviations Proportional to the Mean” in the [Gerow Sample Size and Power Calculator](#). Frequently a better reason is simply that for BIOLOGICAL datasets, variation (sometimes on the scale of SD, sometimes on the scale of variance) is often proportional to means. When one has enough data, one should determine proportionality of the SD vs. the mean and choose options accordingly.

Variations on the theme: The [Gerow Calculator](#) allows one to easily play “what if” games with the options. If only one thing is varied from the above, we can learn the following:

If the only change from the above step-by-step is choosing “Standard Deviations Proportional to the Mean)” rather than “Constant SD” (step 6), the sample size required jumps from 45 to 73 to get the same 90% power.

If the only change from the above step –by-step is choosing “Variances Proportional to the Mean” (also sometimes the case), the sample size required jumps even a bit higher, to 77 to get 90% power. A conservative-safe choice is to try several of these options if one is not too sure which one fits the best (say because sample sizes are small and one is not too sure if the SD or variance is proportional to the mean, and then pick the highest estimated sample size derived from the various options.

What if sample sizes are known or suspected to end up unequal? If the only change from the above step-by-step is choosing “independent, unequal sample sizes” rather than “independent, equal sample sizes” but SDs are still equal, and if we choose sample size of 45 for the first sample to be comparable to the Zar, case, we see that if the second sample was 30 rather than 45, power would decrease to 83%. To get a power of 90%, one has to toggle sample two up to a sample size of 45. These types of calculations can help with “[completeness](#)” goals. In other words, if one

needs 45 samples for power, one often needs to aim for more than that (by 10% or more) knowing or suspecting that some will fail due to various real-world field factors (extreme weather, instrument losses, etc.) and/or some [QC](#) failures that would tend to make some of the data unusable.

Using the [McBride Detection Calculator](#) to Get Zar's Answer: Use Zar equation 8.22. Here is the step-by-step method to replicate the Zar answer of a sample size of 45:

Choose the **Point-null** option, (which is always [two-sided](#)), then choose the **two-group** option, click on n to solve for sample size, then:

- 1) Type in alpha as 5 (corresponds to 5% or a significance level of 0.05), then
- 2) Type in detection probability as 90 (for 90% power),
- 3) Type in effect size magnitude as 69.3 (corresponds to 69.3% of one standard deviation sample = effect size expressed as a % of the true Standard Deviation). Details on how we got to 69.3:

Since effect sizes in the [McBride calculator](#) are expressed as percentages of the magnitude of the standard deviation (SD), to get from the effect size in original units (0.5) to effect size to effect size a % of the standard deviation, we first convert the effect size from 05 in the Zar example to a SD percentage with the equation $\text{effect size} = 100 * (1.0 / \text{standard deviation})$. $\text{MDD} / \text{SD} = 0.5 / 0.721 = 0.693 = \text{ES}$ as a fraction of the SD. How we got 0.721: Zar gave the variance as 0.52, so the SD = the square root of 0.52 = 0.721.

- 4) Then click on calculate and to be precautionary, choose the larger of the two values (45, the same answer Zar gave), for variance unknown. Although we have a rough estimate of the variance (0.52), the true population variance is seldom truly known with great accuracy, so choose the higher sample size to be precautionary.

Sample Sizes Needed for Differences between Two Means When Using Paired Sampling

This section describes how to compute the minimum sample size needed to detect a pre-determined magnitude of difference (minimum detectable difference = [MDD](#)) between two means when PAIRED SAMPLING is being used. One advantage of paired sampling is that one can often get a higher amount of power at lower sample sizes than for sampling that is re-randomized each year.

Often the paired sampling options make sense for comparing different time periods at judgmentally picked sites, where one is taking repeated samples at one site year after year. Comparisons could be for samples from one year to samples the next year

if sample size is large enough (use the paired versions of the calculators to see if sample size is large enough).

Scenario 1: The samples have equal (or similar) variances, SDs, and sample sizes:

For this scenario one can use Zar's example (the step-by-step is just above) and make all inputs to the [Gerow Calculator](#) "constant SD" option exactly the same except for step 9 and 10. Do the following for steps 9 and 10:

Move the arrows in the spin button for "Design Choices" until "paired sampling" appears, then just below that

Move the Sample Size Slider with the arrows until the desired power (90%) appears in the power answer above the Design Choices Box. In this case, a sample size of 28 is seen correspond to 90% power, so the required sample size answer is 28. The smaller sample size of 28 is needed instead of 45 when the only input change is to choose "paired sampling" rather than "independent, equal sample sizes." Planners often decide to do paired samples (which are relevant to non-probabilistic-judgmental designs where one is going back to a certain site each year). As can be seen in our example, paired samples typically have the advantage of requiring fewer samples (providing more statistical power for the same sample size).

However, keep in mind that paired samples are not a "cure-all." Using paired samples does not insulate one from problems related to small sample sizes or asymmetric distributions. "Power decreases as the variance increases, decreases as the significance level is decreased (i.e., as the test is made more stringent), and increases as the sample size increases. A very small sample from a population of paired differences with a mean very different from 0 may not result in a significant t test statistic unless the variance of the paired differences is small. If a statistical significance test with small sample sizes produces a surprisingly non-significant [P value](#), then a lack of power may be the reason. The best time to avoid such problems is in the design stage, when appropriate minimum sample sizes can be determined, perhaps in consultation with a statistician, before data collection begins" (Northwestern University Discussion of [Do your data violate paired t test assumptions?](#)).

Both the McBride sample size calculator and the [Gerow calculator](#) give one the option of estimating sample sizes for paired samples, either when variances are or are not equal.

Using units expressed as % of the true population standard deviation (rather than original units of measure) one can also compare sample sizes needed and statistical power using the McBride equivalence detection probability calculator. The equivalence option needs to be initially chosen so that one can eventually get to the [inequivalence](#) option. However, output answers are given in a format which allows easy comparisons

for power produced at given sample sizes for three cases: 1) equivalence testing, 2) inequivalence testing, and 3) standard null hypothesis significance testing. Paired vs. not-paired options can be selected as part of the first input prompts.

Other tips on using the paired version of the calculator (Ken Gerow, University of Wyoming Statistics Department, Personal Communication 2007):

If you don't know whether or not the mean is proportional to the variances, one can often choose the affirmative option (mean is proportional to the variances) as a first choice. Then just enter in the observed mean, and the ratio (mean/variance) as an estimate of the proportional relationship. Since you are not yet sure of the relationship, however, bracket the estimates of sample size with the answer from another option. In other words, also try the option for "proportional SD" as another flavor of "variation changing with means." With no evidence (i.e. not enough data) with which to choose between proportional variances and proportional SDs, trying both and (perhaps) choosing the one that yields conservative sample sizes (i.e. larger) would be a "safety first" approach.

Sample Sizes for Two Samples, Variances Unequal

It would not be unusual for a standard deviation (SD) to be different in one time period versus a later time period or between different sites. If one has unequal variances, one will also have unequal SDs and one can use the [Gerow calculator](#). One simply chooses "different SDs" in the variation choice box, and then types in the SDs for each sample in the field provided.

The Gerow calculator is a free MS-Excel macro [right click the sample size/power calculator box to download the .xls (Excel format) macro file to your computer] and then choose "save as" to save the file to a location of your choice in your computer. The Gerow macro includes options for the following (often-relevant) scenarios: 1) paired samples, 2) either SD or variance proportional to the mean, 3) for different sample sizes, and 4) either equal or unequal variances. This more-advanced calculator includes instructive example plots in help screens. Be sure to read the help screens carefully as the input variable choices are not quite as quickly understood as some of the other calculators mentioned above. An advantage of Gerow's calculator is the extra options. In real-world aquatic biology datasets, standard deviations (SDs) are often proportional to the mean or variance. The Gerow calculator provides a way to easily draw graphs on the SD/mean and variance/mean comparisons to get hints about which may be true. However, if still not sure (perhaps sample size is just too small to establish good relationships, Gerow suggests that one could calculate the sample size with the multiple options and then choose the largest sample size answer to be conservative. In the real world where some samples fail, sample sizes are often not equal, and the [Gerow calculator](#) also has inputs for sample size not being equal if one chooses independent samples (not paired) in the design choices box.

More information on the Gerow calculator and choices therein are found in a free [download discussion paper](#) associated with the calculator. Among other things, the discussion paper explains why choosing a small sample size for the first year of data collection can greatly limit statistical power that can be achieved in subsequent power

calculations. Bottom line: “This implies, as a matter of practicality, that one ought to do as much as possible in the first year of a monitoring effort to minimize the chance of this occurring”

Other internet sample size/power calculators also cover unequal variances, including a [University of Iowa calculator](#).

Again, it is important to consult a professional statistician before finalizing sampling designs. Although the calculators listed above would often be fine as a first cut (illustrating which variables or strata that are just way too variable to allow networks to detect a difference of concern), all such simplified calculators should be used mostly as a first step. As more data is collected and monitoring plans began to be refined, networks should consult with a professional APPLIED statistician for more advanced fine tuning of samples size and power estimations. Those with more advanced expertise may choose to use more advanced methods (such as the [Gerow calculator](#) or simulation approaches).

To Detect a Stated Difference between a Mean and a Standard

This section describes relatively simple methods to compute a minimum sample size needed to enable one to be able to detect a pre-determined magnitude of difference (think of it as a minimum detectable difference = [MDD](#)) between a mean and a single value, such as a water quality standard or other benchmark. Parks and monitoring networks should check with States first, as sometimes the state specifies how many samples need to be averaged. A geometric mean is specified in some cases for bacteria. If the State has defined requirements (minimum number of samples, timing of samples, type of calculation, etc.) then that should be used and any calculations covered in this section would be considered optional.

To be precautionary in ensuring resource protection (in addition to performing any State-required estimates), Parks may also wish to know if an average statistically exceeds a standard or criterion. In that scenario, one would want to make sure the samples were representative of either typical cases (for comparisons with chronic standards), or worst-case times of day and/or times of year (for comparison with acute standards at the [diel](#) or seasonal change time periods most apt to exceed standards). Once one has determined how to get representative samples for the question of concern, one can use the following to initially estimate the number of samples needed to determine if an average value exceeds a water quality standard, criterion, or other benchmark of concern.

Although there has been some controversy in using one-sided test in medical trials (where one might want to know about both positive and negative effects), such tests are less controversial in helping to answer inherently one-sided questions such as: Does the average value exceed a water quality standard? ([Using Statistical Methods for Water Quality Management: Issues](#), see discussion on page 80). For those who decide they want to use a two-sided test, Zar [1999](#) has equations and example right answers. For similar inputs to those given in the one-sided example below, the required sample sizes may be a bit higher for the two-sided test. For example, using Zar equation 7.8 and one-sample, normal-distribution, **two-sided**, sample size calculators, Zar gives a sample size of 19 for his example 7.7 (page 107) an example which uses many of the same input variables used

just below for the **one-sided** but otherwise analogous case to get a required sample size of 15.

However, many only want to know the answer to the one-sided question (does the value exceed a standard?) and EPA has a [freeware beta test sample size and sample frequency estimator](#) that can be used to estimate the sample size needed to test the difference between a single mean and a specified value (such as a water quality standard). The equation used is sample size = variance times $(t_{\alpha,v} \text{ plus } t_{b(1)v})^2$, all divided by the MDD. For explanation see EPA [documentation file](#). In that calculator, if one inputs the following, variance 1.568, alpha = 5, beta = 90, and MDD = 1.0, the answer returned is “An estimated 15 samples must be collected to yield a 90% chance of detecting a difference as small as 1 at a 95% level of confidence. The estimate is based on a variance estimate of 1.568, a one-tailed 0.05 level of significance (alpha) and a one-tailed beta of 0.1. Corresponding t-values used to calculate the estimate are 1.761 (for $t_{\alpha,v}$, the critical one-sided t-value at alpha = 0.05 at sample size 15, DF=14) and 1.345 (for $t_{b(1)v}$ = critical one-sided t-value at beta = 0.10, or 90% probability, at sample size 15, DF=14) respectively.”

One can get the same answer, 15, using the McBride Calculator with the following step by step:

Choose the **one-sided, one-group** option, and click on n to solve for sample size, then:

- 1) Type in alpha as 5 (corresponds to 5% or a significance level of 0.05), then
- 2) Type in detection probability as 90 (corresponds to 90%), then
- 3) Type in effect size magnitude as 79.9 (corresponds to 79.9% of one standard deviation of the values in single sample = effect size expressed as a % of the true Standard Deviation), the
- 4) Click on calculate, view answer as 15 (choose 15 rather than 14 since we seldom know exact variance but instead usually have rough estimates).

In the McBride Calculator, the part that is not quite as straight-forward is how to get the effect size magnitude as a percent (79.9). Here is the step-by-step: Since effect sizes in the [McBride calculator](#) are expressed as percentages of the magnitude of the standard deviation (SD), to get from the effect size in original units (1.0) to effect size to effect size a % of the standard deviation, we first convert the effect size from 1.0 (original units) in the example to a SD percentage with the equation effect size = $100 * (1.0 / \text{standard deviation})$. In the calculator, the $*(1.0 / \text{standard deviation})$ part is expressed as $\Delta = \frac{\mu - \mu_0}{\sigma} = \text{effect size} = [(\mu - \mu_0) / \text{SD}] = [(\hat{i} - i_0) / \hat{\sigma}]$ for the one-group case. Since Variance is 1.5682, $\text{SD} = \sqrt{1.5682} = 1.252278$. $\text{MDD} / \text{SD} = 1 / 1.252 = 0.7985 = \text{ES}$ as a fraction of the SD. Round that value to 3 significant figures and express it as a % (79.9), the input value needed for the McBride calculator as an effect size.

A key point is that this type of “effect size” is a difference (a MDD in the discussions above) between a mean and a standard MDD value (in original units

of measure) divided by the true SD (in original units of measure). In the McBride calculator, the value 79.9 is typed into the calculator without the % symbol.

Those who are planning to use [Upper Confidence Intervals \(UCIs\)](#) can use the sample size estimators given just above as a first estimate of needed sample sizes. However, in deciding what standard deviations to use in the calculations, keep in mind the following: Among the recommendations in EPA's guidance for assessing contaminated soil at Superfund site (EPA 2002, [Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites](#), EPA Publication EPA 540-R-01-003):

To be precautionary when sample sizes of past or pilot data is small or questionable, before estimating needed sample sizes, "it is advisable to use an 80 or 90 percent upper confidence limit for the estimate of the standard deviation rather than an unbiased estimate to avoid underestimating the true variability."

Solving for Minimum Detectable Difference Rather than Sample Size:

If one already has a proposed sample size, one can also use Zar's (rearranged) equation to solve for [MDDs](#): Zar's minimum detectable difference (equation 7.9, Zar op. cit.) for one sample vs. a water quality standard as follows: $MDD = \text{square root of } [\text{sample variance}/n] * (\text{two-tailed critical t-value for } 1-\alpha \text{ given } n-1 \text{ degrees of freedom and probability chosen} + \text{upper, one-tailed t-value for } 1-\beta \text{ given } n-1 \text{ degrees of freedom and probability chosen})^2$. Again, when using Internet calculators, be careful how the inputs are made as different Internet calculators have different input formats (for example does one input 90% power as 0.9 or as 0.1).

Note: t-value terminology varies in different statistical text books and can be confusing. Herein, the phrase two-tailed critical t-value is synonymous with [two-sided](#) critical t-value. Some authors leave the word critical out of the terminology. Still others use subscripts. One of the more clear explanations of these terms and subscripts is found, as well as a table for both two and one-tailed cases is in Zar ([1999](#)).

Inequivalence (Bio-inequivalence) Sample Size Calculations

In an [inequivalence hypothesis test](#), the hypothesis to be tested is that a difference between population means lies beyond a stated interval.

One can use the McBride detection probability calculator not only to reproduce some of the same Zar-example values for the null hypothesis test obtained with the [Gerow calculator](#), but also to compare probability of detection (or required sample sizes) for inequivalence tests (always precautionary), equivalence test (for the most part never precautionary enough for common NPS purposes), or a standard null hypothesis t-test (often very precautionary at high sample sizes but decidedly not precautionary at low sample sizes).

It is hard to determine if null hypothesis tests (NHST) are precautionary or not if power is not controlled and specified, but usually NHSTs are not as precautionary as inequivalence testing at low (especially at less than 30) sample sizes.

Be aware that using alpha of 0.05, the use of inequivalence tests typically requires much larger sample sizes than the use of equivalence tests, and also larger sample sizes than standard null-hypothesis tests. These factors can be especially important when one is otherwise initially tempted to consider very low sample sizes as a starting point.

For smaller sample sizes, we would ordinarily favor inequivalence testing to equivalence testing (or to null hypothesis significance testing) in the NPS, to take a more precautionary stance in limiting conservationist (type II) risk. After all, the NPS is charged with protecting rare and valued resources for future generations. Therefore, whether journal editors understand the concept or not, it is illogical to limit the polluter's risk (alpha) to lower levels than the conservationist's risk (beta).

When considering these issues, it is helpful to reconsider the questions to be answered, the objective of the testing. The key point is that in a precautionary approach—testing the inequivalence hypothesis—large sample sizes may be needed in order to conclude that some effect actually is *not* environmentally important. With small sample sizes, it will be hard to prove that

Although standard null hypothesis tests have come under fire in recent years, they can be used as one line of evidence. This is not a bad idea when sample sizes are large, say over 30-50, depending on variability magnitudes, as long as one also states the effect size or minimum magnitude of the difference to be detected AND as long as beta is controlled to small levels (0.01 or 0.05). The past too common practice of just specifying significance level (alpha) alone and leaving beta unknown and uncontrolled is not recommended for standard null hypothesis significance testing relating to valued NPS resources.

In general, proof of safety (as in an inequivalence test) is harder to establish than proof of hazard. Figure 5.11 in the [McBride statistics book](#) shows that the point-null test can behave very differently from an inequivalence (or equivalence) test. In inequivalence tests, to reject the hypothesis is to infer that the variables being sampled are very unlikely to differ by more than a specified amount (the prescribed interval width) and so may be considered as equivalent. That is, differences are expected, but if they are small enough the variables can be considered as bioequivalent.

One can sample size what-if games with the precautionary inequivalence option vs. other options given various effect sizes and sample sizes (and various alpha and beta inputs) using the McBride Detection Calculator to see how detection probabilities change under different scenarios.

The [McBride calculator](#) is in units expressed as a % of the standard deviation, but for comparison, one can do paired sampled calculations **in original units** of measure for a standard hypothesis test using the [Gerow Calculator](#). Using the McBride calculator, be aware that in a perfectly normal population a standard deviation might be as much as three times smaller than either of the means being compared, so an effect size as a difference between means of 50% of the standard deviation (considered a moderate effect size in high sample size psychological studies), might be considered small when compared to the mean in smaller (and often very skewed) data sets more typical of water quality or aquatic ecology. Unless one was very close to a resource collapse [threshold](#)

and/or sample size was large and variance was small, most of those monitoring outdoor environmental variables would not even try to detect such a small difference as the magnitude of 50% of the standard deviation over relatively short periods of time, say a year or two.

In typical (skewed) field environmental data sets, a standard deviation can be half or even equal (or greater) than the magnitude of the mean, so after using the [McBride calculator](#), translate the values back to percentages of the mean or median to get a reality-check look at the effect size from a different angle. The equation used to translate back to original measurement units is effect size = the difference in means or the difference between a mean and a comparison water quality standard in original units of measure = the effect size (as a percent of the standard deviation) times the standard deviation, all divided by 100.

The factors that have retarded the use of inequivalence testing have included:

1. Many are not yet familiar with it.
2. Sample size requirements for inequivalence testing can seem high. However, low sample sizes are potential factors in a plethora of statistical pitfalls, so assuring reasonable sample sizes is not a bad idea. In fact, one advantage of considering inequivalence testing is that when one determines detection probabilities, it will steer one way from the very low sample sizes (often less than 25-50) that can be so problematic in standard null hypothesis significance testing.
3. There are not as many statistical packages include options for it.
4. There are few if any user-friendly sample size calculators available based on original units of measure. But one can always use the [McBride calculator](#) (op. cit.) to estimate needed sample sizes after converting minimum detectable differences in original units to percentages of the standard deviation.

Comparing Inequivalence, Equivalence, and NHST Options:

The following can be used as a step-by-step procedure for comparing the sample sizes needed for the three different options:

Equivalence Testing Versus Inequivalence Testing

For consistency with an example already given in the discussions above (in the step-by-step sections above for both Gerow and [McBride Calculators](#) for a Difference between two means), let's use the same Zar example input variables for equation Zar equation 8.22 (Zar, [1999](#); significance level = 0.05, ES = 69, and sample size for each sample 45). At the McBride detection probability home page:

- Choose equivalence hypothesis and two groups, the click on “D” for detection probability. This is the only choice, but be aware you have to choose it even though you are going for inequivalence testing rather than equivalence testing.

- Type in 5 (for the 0.05 significance level), then choose no for paired data and choose next.
- Choose parallel (the common choice), then choose next.
- Choose ES, then choose next.
- Type in 69 (the ES to the nearest integer) for DeltaU, then type in 100 for Delta Max, then type in 45 for the sample size of each group.

Next, read the probability of detection on the first line opposite the -100 figure. For the (default-choice) TOST inequivalence test option, the detection probability is 99.9%. For the (default-choice) McBride equivalence test option, the detection probability is 46.2%. In other words, using inequivalence testing, there is a 99.9% probability that that the tested hypothesis (inequivalence) will not be rejected. For the equivalence testing option, the conclusion is that there is only a 46.2% probability that that the tested hypothesis (equivalence) **will** be rejected. Obviously the inequivalence test is the more precautionary choice and more appropriate choice for are resource protection/stewardship agency than an equivalence test. The fact that it is more precautionary is the very reason that inequivalence testing is used in drug testing, one cannot afford to make a mistake that would harm humans. Since the NPS tends to be precautionary about protecting endangered species and other rare trust resources, the NPS would ordinarily use the more precautionary inequivalence testing rather than equivalence testing.

Inequivalence Testing Versus NHST Testing

Choosing between inequivalence testing and the more familiar null hypothesis significance testing option (NHST), using the same input variables used just above: We already know (and could independently confirm using the [McBride calculator](#) null hypothesis, two sample option), that the detection probability for the NHST option for the example input variables used just above would be 90%. So in this particular example (sample size 45), the NHST option would be fine to assure 90% power in detecting the stated magnitude of difference (1.5 in original units in Zar's example 8.4, page 134, Zar [1999](#)), as would the inequivalence test option. To assure 99.9% detection probability (analogous to power) would take a higher sample size for the NHST option, giving an advantage to the inequivalence testing option. However, the comparison above clearly shows that the equivalence testing option at sample size 45 is clearly not precautionary enough for NPS purposes, with a detection probability of only 46.2%.

Bottom line: the NHST will often be precautionary "enough" for NPS purposes at higher sample sizes (usually above 25-45), but not at lower sample sizes. At lower sample sizes, the inequivalence hypothesis test should be used as the first default-choice rather than the NHST. One advantage of doing this is that the inequivalence option will instructively and clearly show the lower detection probabilities (and thus the need for higher sample sizes) of the NHST at low sample sizes.

Sample Size Needed to Estimate a Single Proportion

Ratios, proportions, and percentages are intuitively appealing and understandable ways to express relationships between two variables. Many (such as ratios between medians) are appealing for biological relationships.

However, calculating single proportions well (with a reasonably small confidence interval about the proportion) is notoriously difficult and usually should not be attempted with a sample size less than 25-50. If one has fewer samples than that, and there is no way to remedy the small sample size, one should probably be transparent that the proportion estimate is probably not very accurate (the confidence interval about the proportion is too wide to claim reasonable accuracy).

Calculating the exact needed sample sizes for the estimation of a proportion is typically done in EMAP style surveys. The proportion of stream length or miles impaired is an objective of interest to some parks and monitoring networks. Part of the appeal is that it can relate to GPRA and other more general goals (desired conditions, condition assessment percentiles, ecological [thresholds](#), etc.). One typically should use probabilistic monitoring designs for questions such as: “What percent of stream miles are impaired?”

One can make initial estimates of needed sample sizes with table 1 and the equation in EPA’s discussion of “[How many sample sites to use?](#)” In this case, sample size calculations depend only on the proportion and desired % confidence (a z-distribution confidence interval on the proportion) required. Note that EMAP is using the word “precision” in a nonstandard (compared to the more normal QA/QC or [NIST/ISO](#) terminology for [precision](#)) way here. What EMAP (and many statisticians, for that matter) means by precision is a z-distribution confidence interval surrounding a summary statistic (a proportion in this case), rather than measurement precision. The *t*-distribution that is used for smaller sample sizes for means can’t be used for proportions because it is not applicable to sampling from a binomial distribution (the same is not true for the z-distribution at larger sample sizes). The confidence interval equation should only be used at sample size 25 or above ($n = 50$ is the recommended default) but does not depend on a normal distribution.

For more details, including a step-by-step example for using the equation, see [Part B](#). A minimum sample size of 25 relates to different sites rather than to replicate times of day at the same site. However, same-site temporal variability will have considered when making data analysis decisions, to help make sense of the data related to standards exceedances. A key common sense question is how much change in time or conditions can occur before the sample is no longer one sample (and maybe sample size is no longer big enough to estimate a proportion reasonably well).

EPA has a [freeware beta test sample size calculator for estimating a single proportion](#) with various degrees of confidence (size of confidence interval). The equation used is sample size = Z value [the $a(2)$ version, which would be 1.96 if significance is 0.05] squared times the initial estimate of the proportion (for example use 0.5 for 50%) all times q (where $q = 1 - p$, where p is an initial estimate of the proportion), then all divided by the length of the confidence interval expressed as a % of the proportion (use 0.2 for 20%).

If one does not know what a single proportion will be before sampling, use 50% as a starting point in the calculator to be precautionary (to make sure sample size is large

enough). A trial illustrated that the EPA guidance for starting with a sample size of 25-50 for a proportion is in the right ball park, as with a proportion of 1/2 (50%) the EPA Z calculator returns the following answer: “Based on an initial guesstimate of the proportion as 50% (worst case, since 50% requires the largest sample size), an estimated 24 samples must be collected to estimate the proportion within plus or minus 20%, at a 95% level of confidence.” If a confidence of a plus or minus 10% is needed, a larger sample size (96) is needed than for 20%. The beginning estimate of the proportion must be between 0.1 and 0.9, and changes in the magnitude of the beginning estimate of the proportion will impact the result in sample size. For additional discussions and examples for various assumed proportions see EPA discussion of “[Why a sample size of 50?](#)”

Using simplistic human-population-based calculators (such as the University of Connecticut [confidence interval sample size calculator](#)s) designed for normal populations and large sample sizes to estimate confidence intervals about a proportion would seldom be optimal in outdoor environment monitoring and using them could result in misleading conclusions. For one thing, the sample size of the [Target Population](#) of real interest is typically hard to quantify but is usually very large. Often the true [Target Population](#) might relate more to all the samples that could be obtained if one were sampling much more frequently (every 5 minutes over a one year period, 24-hours a day, for example). Assume the target population had to do with four rivers. Would the size of the target population of interest (to plug into the U. of Connecticut calculator, just above), then ever simply be 4? No, that is not really the target population of interest. What is more relevant to biota trying to survive in the rivers is the changing magnitudes of things like nitrogen concentrations, temperature, and pH, all of which can change drastically in one 24 hour period, even more seasonally (often a big difference from January to August), and year to year (sometimes a big difference from one year to the next). So the target population of interest is hardly ever 4 rivers, say sampled twice in 5 years, and the sample size of the target population of interest to resource managers would basically never be 4 in that case

A more common use of calculators relevant to proportions would be to say there are 4 rivers in the park, which we are going to sample randomly each year. How many samples do we need to estimate the proportions of total riverine sites in the park considered impaired each year? Assuming the true proportion is close to 50%, looking at Table 1 at the [EMAP discussion](#), one can see that one can estimate the proportion with a confidence interval of a plus or minus 20% about the estimated proportion with 95% confidence with 25 samples, or one can estimate the proportion with a confidence interval length of plus or minus 10% about the estimated proportion with 95% confidence with 100 samples.

Although trying to get 30-50 samples is a good rule of thumb to help prevent collecting **far** too few samples when estimating accurate proportions, if an event is rare enough (which might be true in pristine areas of National Parks), even 50 or more samples may not be large enough to have a very accurate (very small confidence interval about the proportion) estimate of a proportion. For proportions, larger sample sizes lead to smaller standard errors (smaller confidence intervals about the proportion), and the relationship is not dependent on standard deviations.

Perhaps unwittingly further supporting the “dangerous equation” notion in the title, a recent paper that seems to use terminology a bit too loosely and move from one

topic to another without good transitions, and thus tends to confuse or not optimally discuss the differences between:

Means and Proportions and SEs about means vs. SEs about proportions,

Sample SDs with SEs about means (the latter is a type of SD but not a sample SD, so one needs to be careful with terminology)

Simple variation (as say a sample SD) and variances (squared SDs),

A good example of the hazards on estimating proportions when sample sizes were relatively small versus **very** rare events.

When looking at national health statistics, at first glance it may appear to the reader that very rural counties with very small populations have lower proportions of the citizens with a rare kidney cancer than more populous counties. However, the standard error of the distribution of proportions is fully dependent on the sample size, so small counties have less accurate (larger sample size) estimates of the true proportion than large counties. A county with 100 inhabitants that has no cancer deaths would be in the lowest category of proportions of citizens with the affliction (0/100). But if that same county just so happened to have had one cancer death instead of zero, it would then have a reported proportion of 1/100 and would then have been lumped among the counties with the highest rates of the disease. Counties like Los Angeles, Cook or Miami-Dade with millions of inhabitants do not bounce around like that (Wainer, H. 2007. [The Most Dangerous Equation](#). American Scientist 95:249-256).

Sample Sizes to Estimate a DIFFERENCE between Two Proportions

The EPA beta test sample size calculator has Z distribution options for calculating sample sizes needed for (see EPA beta test sample size estimator [Version 0.7.2.2](#)):

1. A test for a difference between proportions and
2. A test for a difference between a proportions and a value.

If the unzip and install of the beta test is not compatible with your version of windows, contact the author for alternative ways to install the beta version.

Cases Where Variances Are Not Calculated in the Usual Way

Remember that for status assessments sometimes one doesn't need a sample variance: If sites are chosen probabilistically rather than judgmentally, and the questions are more along the lines of "what % of park river miles are impaired," then paired designs or estimations of sample sizes based on paired comparisons are not relevant, nor is variability a big player in the estimations of needed sample size. For most questions to answered in terms of proportions or percentages (% impaired, for example), networks can

usually just plan on the need to obtain at least 25-50 samples in the sampling period of interest (see EPA discussion of minimum detectable differences in proportions and “[why a sample size of 50](#)”, and EPA general [sample size discussions](#)). When the need is to find sample sizes, the sample variance does not play into the equations.

However, there are other cases where one needs a variance, and the best way to calculate the variance is not always readily apparent. This is especially true for complex (unequal probability of selection) monitoring designs. Regardless of the design chosen, good estimates of variance or standard deviations are needed for all of the more rigorous sample size calculations. Again, beware of cases where the full range of conditions is not covered, or the sample size is very small.

Complex Variance Methods for **Status Assessments**

The goal is to get an accurate one-time (status) estimate of the variance based on unequal probability of selection designs (such as [GRTS](#)). One “different way than normal” method to estimate variance for GRTS status assessments is the “local variance estimator” (sometimes also called the “local neighborhood variance estimate”). Calculating this requires complex equations. Part of the goal of GRTS users in separating “local variance” from other contributors to total variance, seems to be to get the variance down, but one cannot thus get rid of some contributors (such as lack of perfect measurement [precision](#) of repeat measures of a single homogeneous sample).

If networks decide to pursue the local variance estimator approach, they should keep in mind that these approaches are complex enough to often require the help of knowledgeable statisticians.

[Complex methods to estimate variances for trends](#) are discussed separately below.

For examples of calculating variance for status estimates in complex ways and related discussions, see:

Stevens Jr., D. L., and A. R. Olsen. 2003. Variance estimation for spatially balanced samples of environmental resources. *Environmetrics* 14:593-610.

Stevens, D. L., Jr. and A. R. Olsen 2004. Spatially-balanced sampling of natural resources. *Journal of American Statistical Association* 99(465): 262-278

However, keep in mind that the very complex ways to estimate variance may not always be needed. For example, in conclusions of one recent paper, simulations suggested that the "naive" (standard way to estimate) sample variance should work well despite being design biased, except when there was a high correlation between the response and the auxiliary variable. However, in multipurpose environmental surveys, this type of high correlation is unlikely to occur [J. Courbois and N. Urquhart, 2004. Comparison of Survey Estimates of the Finite Population Variance. *Journal of Agricultural, Biological, and Environmental Statistics* 9 (2): web-accessible [Abstract](#)].

An alternative to the local variance estimator is variance based on Horvitz-Thompson variance formulas (requires approximation to joint inclusion probabilities). Both of the Horvitz-Thompson Variance Option and the Local Variance Estimator option are included as options in the EPA [“spsurvey” downloads](#).

However one makes the calculation, once an optimal variance estimate is decided, then networks can take the square root of the pooled sample variance to get a (type of pooled) sample standard deviation to plug into some of the sample size calculators discussed elsewhere herein. The simple power and sample size calculation results can then be roughly compared to results from more elaborate methods.

The simple sample size and power calculations can then be compared with more exotic varieties.

The following references are ones that Tony Olsen of EPA EMAP has found to be useful in that they incorporate additional information that impacts sample size calculations. Cochran is the simple usual approach, Harris incorporates a tolerance coverage concept, and Guenther (for means) adds uncertainty in standard deviation used in the calculation. The eventual plan is to include sample size calculation script that can run in R in our `spsurvey` (op.cit.) library of functions for survey design and analysis (Tony Olsen, EPA EMAP, Personal Communication, 2007):

Cochran, W. G. 1987. *Sampling Techniques*. 3rd edition. John Wiley & Sons, New York.

Greenland, S. 1988. On sample-size and power calculations for studies using confidence intervals. *American Journal of Epidemiology* 128: 231-237.

Guenther, W. C. 1973. Determination of sample size for tests concerning means and variances of normal distributions. *Statistical Neerlandica* 27:103-113.

Harris, M., D. G. Horvitz, and A. M. Mood. 1948. On the determination of sample sizes in designing experiments. *Journal of the American Statistical Association* 43:391-402.

Kupper, L. L., and K. B. Hafner. 1989. How appropriate are popular sample size formulas? *The American Statistician* 43:101-105.

Composite Samples, a Special Case

How does one determine statistical power in relationship to sample sizes when compositing many individual fish into single composite tissue samples for contaminants analyses? In 2000, EPA provided look-up statistical power tables that illustrate that as “a factor similar (sic) to a coefficient of variation (CV)...as the ratio of the estimated population standard deviation to a screening value (SV) increases (i.e. SD/SV), the statistical power decreases” (see EPA 2000. *Guidance for Assessing Chemical Contaminant Data for Use in Fish Advisories*, Volume 1, ([section 6.1.2.7.1](#)). Although published in 2000, EPA was still recommending this same publication (as well as probabilistic sampling of fish tissues) in 2006 (EPA, 2006 [Draft Guidance for Implementing the January 2001 Methylmercury Water Quality Criterion](#), EPA-823-B-04-001). In that same document EPA recommended: “To address spatial variability of methyl mercury levels in fish, EPA recommends that states and tribes design a probabilistic sampling by randomly selecting sites or sampling locations.”

Don't put all your eggs in composite samples at first. Although composite sampling provides good estimates of mean concentrations within a species at a location, the true variability, maximum concentration, and the spatial distribution is lost. In composite sampling, it is therefore notably helpful to understand differences in means and variances (between compositing and not compositing) on a few pilot-scale (trial) samples before finalizing composite sampling schemes for very large and expensive monitoring projects. As mentioned in the next section on bacteria, until one is sure about the variance estimates to be used in sample size estimators as well as the distribution of the parameter of interest in the environment being sampled, one should probably take discrete samples for a while at first rather than immediately jumping to composite sampling. For more details, see [Part B](#).

Gilbert has a whole chapter on compositing, providing complex formulas for estimating the variance of means and required sample sizes (Gilbert, R.O. 1987. *Statistical methods for environmental pollution monitoring*. Van Nostrand Reinhold Co., pages 6 and 72).

In considering composite samples and sub-samples, making sense involves understanding what Cochran calls "relative precision," the ratio of the variance from the combined sample (local small area plus larger area) over the variance from the larger area sample (W.G. Cochran. 1977. *Sampling Techniques*, 3rd edition, John Wiley & Sons, New York)

Bacteria Sampling: Another Special Case

Like pH, bacteria samples are different partly because they are already on the log scale, because of lag times and some other unusual ways in which they are used, and because sample size estimations, power, and compositing are all handled a bit "differently" than for most other parameters. Guidance on all these topics, including the concept that one should not composite at first (until one is sure about the variance estimates to be used in sample size estimators as well as the distribution of bacteria on beaches) was provided by EPA (EPA, 2005, [The Impact Beaches Report](#), Publication EPA 600/R-04/023).

Transects, Another Special Case:

Transects are often used in long term monitoring and they have appeal for various reasons, but power, sample size, and variance estimates can all be more complex than for less complex. For an introduction to statistical aspects, variance estimates, comparison of variance and covariance values, and how spatial correlation complicates the estimates, see Urquhart 2000. [Adapting a Physical Habitat](#).

.Andrea Atkinson and Colleagues of the South Florida/Caribbean network used similar (to those the section above) but different equations derived from Thompson et al. 1998 (W. L. Thompson, G. C. White, And C. Gowan. 1998. *Monitoring Vertebrate Populations*. Academic Press, 365 pages) to estimate needed sample sizes to find a 25% change in mean proportions over 5 years in % cover of living coral, using 20 transects per reef (sample size for each transect to estimate the proportion was >250), where variance

estimates were based on an analysis of covariance (Miller et al. [Lessons Learned During 7 Years of Coral Reef Monitoring](#)).

There can be more information content in proportions based multiple randomly selected transects than in randomly selected single data points, and going back to the same transects each year increases power (but decreases DF). However, these methods are considerably more complex than most of the others discussed herein, and back transformation introduces bias and should not be done without a good justification and/or bias correction. Therefore, before using similar methods, we recommend that all such methods be discussed with a professional, applied statistician to see if they are optimal (vs. simulations and other relatively complex options) after considering assumptions, local factors, and other specifics. The methods that Atkinson used are starting to get beyond the other comparatively simple ones discussed herein that many quantitative ecologists should be able to perform and fully understand without some outside help.

Sample Sizes Needed for Confidence Intervals in General:

Caution: Specify the Kind of Confidence Interval: When using the phrase “confidence interval” (CI), be sure to say exactly what kind. Is the confidence interval to be calculated one or two-sided, parametric or nonparametric, and what summary statistic does it surround? If it is a parametric confidence interval about a single mean, is it t-distribution confidence interval for small sample sizes or a large sample size z-distribution confidence interval about the mean. If the confidence interval relates to means, is it a confidence interval surrounding a single mean or about a difference between two means? Make it clear how the interval will be calculated in the data analysis SOP. If the confidence is one sided, is it the upper or lower confidence interval? If the confidence interval to be calculated will be two-sided what magnitude will be expressed, the whole interval on each side or the (more commonly reported) [half-width](#) about each side of the summary statistic? Making this clear becomes even more important in cases where the magnitude of the confidence interval triggers another decision, such as whether or not the sample size is large enough (see [Include a Cumulative Bias SOP](#) section below).

For sample size estimation, it may be tempting to simply say that “Our sample sizes are driven by budgets only and we are not going to do hypothesis testing, we are just going to estimate confidence intervals and the width of the confidence interval therefore will be our estimate of uncertainty, so therefore we don’t have to do sample size calculations.” There is some truth in this, and the simplicity (and reduced pre-monitoring work load) of this approach has initial appeal. However, for various real-world (and eventual) data interpretation reasons related to long term monitoring, this type of answer is almost always less than fully adequate.

Probably a more common answer for a sample size goal for a confidence interval about a mean is: “Our sample size needs to be big enough to get the confidence interval down to a reasonable size (or to a size consistent with project needs).” So, just as one example, one could do the calculations and then state that a certain sample size is needed to estimate a half-width confidence interval about each side of a mean no larger than $\pm 20\%$ of the mean.

Although this would be better than simply saying we are not going to estimate sample sizes needed at all, typically such answers are not fully sufficient solutions either. Monitoring planners need to consider needed sample sizes in the broader context discussed below.

Although it is true that (strictly speaking) one need not calculate power when no hypothesis tests are planned, this is not the only consideration relevant to long term monitoring. If the long term data are found to be useful, sooner or later someone will want to do other statistical tests with it. This might focus on the difference between one time period and another (which relates to trend detection) or a difference between one sub-region in space or time and another. Even those who plan to do complex time series (repetitive measure) analyses including trend tests often do common sense checks somewhat similar in theory to a t-test or paired t-test as a part of basic functional data analyses, especially as part of the data analyses step following a data summarization step.

When considering confidence intervals it is also good to keep Abelson's caution that "Under the **Law of Diffusion of Idiocy**, every foolish application of significance testing will beget a corresponding foolish practice for confidence limits" ([Abelson 1997](#)).

Relative to needed sample sizes, consider the following:

:

1. Most monitoring networks will eventually want to do tests for trends (such as seasonal Kendall tests for trends), and anytime there is a statistical test, adequacy of sample sizes and power are fair questions. Indeed, some monitoring groups will not "call a trend" unless a trend test indicates and trend AND sample size calculations indicate that sample sizes were adequate to call a trend based on a change a certain minimum magnitude, while assuring a stated degree of statistical power.
2. Likewise, anytime someone proposes a monitoring plan to detect long term trends, one reasonable question a resource manager might have would be: "what is the minimum detectable difference that your sampling design will be able to detect, and over what time period?" To answer the question, one typically needs to assess the adequacy of sample sizes with some attention to statistical power or other relevant detectability issues. Too often, not trying to estimate needed sample sizes and power aspects before monitoring began has been one reason why much past water quality data has been practically unusable for trend analyses.
3. A frequent goal for Vital Signs monitoring is whether or not a resource collapse threshold or other benchmark (water quality standard, etc.) has been passed. Deciding how confident one is about this at least indirectly requires attention to sample sizes and minimum detectable differences. Confidence intervals about **differences between** a mean and a water quality standard, between a mean and a collapse [threshold](#), or between a mean and other benchmarks become exactly analogous to performing a hypothesis test! (Graham McBride, NIWA, New Zealand, Personal Communication to Roy Irwin, 2007).
4. By simply saying we are not estimating power or minimum detectable differences, since we are only calculating confidence intervals, a monitoring network might tend to less homework on identifying

parameters or strata that are so variable that one would be unlikely to find even a large trend difference over even a very long period of time. Again, the result might be generating data not useful in answering practical questions, and missing the need identify and then throw out an impractical measure in favor of a indicators that is more useful for trend detection or resource management.

5. A QA/QC basic is data "[completeness](#)." One needs to determine the minimum sample size needed before one can work backwards from that and state that some % (say 90%) of samples would meet completeness goals.

Sample Sizes for Two-Sided Parametric Confidence Intervals about a Single Mean

Even those networks that don't express any eventual desire to test for trends will typically nevertheless eventually want to ensure that their confidence intervals about the mean are small enough (say a certain % of the mean) to represent a reasonably adequate (i.e. the confidence interval is reasonably small) estimation of the magnitude of the mean. Some statisticians refer to this concept as precision, although it is different than typical [QC or control chart kinds of precision](#).

The ubiquitous t-distribution confidence interval about a single mean is dependent on a calculated standard deviation (SD). In cases where the sample size is very small (and/or the spatial and temporal coverage of the samples are inadequate), the standard deviation has little likelihood of reflecting the full range of conditions of the true but-unknown underlying population SD, and the calculated confidence interval will thus also not be representative of the underlying population. In this case, the width of the confidence interval will not likely be a full accounting of true uncertainty. Usually it is not possible to achieve perfect [precision](#), perfect lack of [bias](#), and perfect [representativeness](#). So even though the formulation of the t-distribution corrects for the flakey SD estimates to some degree, at small sample sizes the sample standard deviation often tends to underestimate the true underlying population standard deviation. Thus, it is not correct to say the % confidence of a t-distribution confidence interval about a mean is a **full** accounting of uncertainty, especially when the confidence interval is based on very small sample sizes.

A key factor to consider is that even a very large starting sample size would not be adequate to give a good estimate of the true but unknown population SD if the samples taken do not cover the full range of conditions and/or if the samples taken happen to be not-representative of the target or population and/or do not give one a reasonable idea of the shape of the population distribution of the underlying target population.

However, possibly an even bigger concern is that whether monitoring planners agree or not, sooner or later, some data user is going to (wrongly) line up some confidence intervals about means to see if they overlap and (sic, therefore are "statistically different") in a pseudo-hypothesis testing manner Many continue to do something like this even though it will give the wrong results. Admittedly, the fact that doing this is invalid is not an especially clear argument about why one should pay attention to sample size when calculating confidence intervals. However, many of the ill-

informed will probably continue to do this, and the conclusions that that they draw from such comparisons will likely be even farther from reality if sample sizes are too small.

CIs about a single mean are perhaps among the most commonly used and are easily calculated. For example, MS Excel easily calculated 95% t-distribution confidence intervals about a mean:

More detail: In the MS Excel Descriptive Statistics popup confidence interval results are labeled simply as "Confidence", and amount to the half-width of the 95% confidence interval: viz., the number that is to be *subtracted* from the sample mean to yield the *left* end of the confidence interval, and is to be *added* to the sample mean to yield the *right* end of the confidence interval ([UT description of Excel Confidence Intervals](#)).

Not so widely understood is the fact that it takes more samples to estimate a mean adequately (confidence interval small enough for intended purposes) than many seem to realize, especially if variability is high and/or distributions are not symmetrical. Ideally one would also have normal distribution, but for real world distributions symmetry is perhaps more important.

Properly calculating the sample size necessary for an optimal parametric confidence interval (CI, such as a *t*-distribution confidence interval) on a mean is complicated and subject to more pitfalls than many seem to realize, especially when very small samples sizes (<10 or 25 depending the data) are involved or when folks are (wrongly) trying to stretch the meaning of a CI (for example with the pseudo-hypothesis test discussed above).

One should probably evaluate normality with probability plots for all data between $10 < n < 25$ to judge the advisability of using a normal theory interval. The smaller the sample size, the less symmetrical the data, the less confidence one has the data is from a normal population (and it is difficult to decide from small data sets). With very small datasets, one is usually also less sure that the values sampled represent the full range of conditions of the [Target Population](#). The more questionable these factors are, the less sure one should be of the validity of calculated parametric confidence intervals. See [Part B](#) for additional rules of thumb and detailed discussions.

Again for emphasis, sample size calculators typically need a good estimate of the SD, and unless variability in time and space is very low, one does not typically have that in very small samples, especially in the skewed data typical of field and lab environmental variables (and the frequent lack of coverage of the full range of conditions in the sampled population compared to the target population). Again, since many (wrongly) use confidence intervals a bit like (pseudo) hypothesis testing, sample size calculators that test for **“a difference between a mean and a single value”** (AND require both alpha and beta inputs) are probably safer ways to calculate sample sizes for CIs about a mean than the most simplistic of the single-sample size calculators for confidence intervals. Better yet, just use confidence intervals the right way.

Any comparison of 1 or 2 observations to a pre-existing group (a pseudo-hypothesis test use) by seeing whether the new observations fall outside an interval built from that group should be done with a prediction interval rather than a confidence interval. A z-interval should never be used in environmental work, as it assumes that

sigma (true but unknown population standard deviation) is known. We never know that. Regarding t-distribution single sample-size calculators, without accounting for power, the t-distribution CI sample-size formula is set at 50% power. So it should be expected that the true interval will be wider than the calculated one 50% of the time. Beta needs to be considered for real-world problems. A good reference on sample size formulas and a sample size estimator for tolerance interval is Kupper and Haffner, 1989. How appropriate are popular sample size formulas? *The American Statistician* 43, p. 101-105 (Dennis Helsel, USGS, Personal Communication, 2006).

Seemingly recognizing that 50% is not good enough, when discussing the inadequacy of the relatively simplistic t and z equations to estimate sample sizes for confidence intervals, Blackwood stated that the simple t and z statistic formulas that specify only alpha and not beta do not give a reasonable degree of confidence that pre-specified confidence interval lengths will actually be as small as specified [Blackwood, L.G. 1991. Assurance levels of standard sample size formulas (*ES&T* 25(8):1366-1367), for more details see [Part B](#)].

Thus, to assure that confidence intervals are based on adequate size to ensure applicability for expanded or potentially implied uses, avoid calculating sample sizes with relatively simple t-value “single sample” confidence interval sample size calculators with no input for beta. An example of such a calculator would be Zar’s equation 7.7 (page 105, Zar, 1999). The equation is variance * [two-tailed](#) critical t-value, all divided by the d squared, where d is the half-width of the desired confidence interval. Zar admits that accuracy of the equation is not very good at small sample sizes, partly because sample variance is not a good estimate of the true but unknown population variance, and that the equation must be solved iteratively with smaller and smaller sample sizes. This is the same equation used to “estimate a single mean” (in the EPA frequency and [sample size calculator](#)). Many (including Helsel, D.R. and R.M. Hirsch. 2002. [Statistical Methods in Water Resources](#). US Geological Survey Techniques of Water Resources Investigations) have explained why these types of simplistic calculators with no input for beta should be avoided and why even those that have input for power should be considered rough estimates for various reasons.

Some might object to calculating power with two-sample calculators when a two sample hypothesis test is envisioned. However, doing so solves some of the problems listed above.

Looking at power does not imply we have to do a hypothesis test, and looking at power is one way to evaluate different monitoring designs including different revisit schedules in panel designs (S. Urquhart 2006, [Designing Surveys over Time](#)).

Given the above realities about how long term data is typically used, planners are advised to pay attention to sample size issues and power aspects. Simplistic sample size calculators for confidence intervals about means that consider only alpha and not beta are not sufficient. Many (including Helsel, D.R. and R.M. Hirsch. 2002. [Statistical Methods in Water Resources](#). US Geological Survey Techniques of Water Resources Investigations) have explained why these types of simplistic calculators with no input for beta should be avoided and why even those that have input for power should be considered rough estimates for various reasons.

When estimating a CI about either side of a mean, be a bit suspect of those based on small sample sizes. If sample is under 30 (and especially under 20), talk to your

statistician, or at least calculate sample sizes multiple ways (with different assumptions, as [listed above](#)) and then adopt the highest answer from the options. Also be a bit suspect of values based on sampling that was not fully representative of the full range of values in the [Target Population](#) in time and space, even when the t-distribution is used instead of the z-distribution.

Although it is true that null hypothesis tests are used as stand-alones less and less in field studies due to well publicized and very real shortcomings (see discussion on choosing alpha, above), in this “anti-hypothesis test climate” there is nevertheless a tendency to simply calculate a series of confidence intervals about means in original units and then (in a pseudo-hypothesis test mode) see if the intervals themselves overlap. However, confidence intervals about means may overlap yet there could be a statistically significant difference between the means, see discussion paper by Bower on Minitab Homepage ([Some Misconceptions about Confidence Intervals](#)).

There are many ways various types of confidence intervals, error bars, and low p-values are frequently misinterpreted, see Di Stefano et.al. 2005 [Di Stefano, J. F. Fidler, and G. Cumming, 2005 [Effect size estimates and confidence intervals: An alternative focus for the presentation and interpretation of ecological data](#), In A. R. Burk (Ed.) (2005). New trends in ecology research. Nova Science:71-102].

Perhaps even less understood is that even overlapping confidence intervals about medians (often expressed as notches in box and whisker plots) are only a crude guide to “significant” differences between medians. Proper hypothesis tests do not look at whether confidence intervals around medians or means overlap or not (Dennis Helsel, USGS, Personal Communication, 2006).

In summary, too often confidence intervals about means (rather than confidence intervals about a difference in means, see next section) are being calculated for implied uses (such as the pseudo-hypothesis tests) by those with minimal knowledge, even though doing so is often not justified in context and most statisticians would say these types of pseudo-hypothesis tests are not a valid replacement for a proper hypothesis test where both alpha and beta are controlled at reported magnitudes.

In this climate, be especially wary of the most simplistic (one) sample size calculators (those that don’t take into account beta), especially for our typically skewed environmental variables. A key question is: How will data, the summary statistics, and confidence intervals, be used, not only now but after long term monitoring has collected enough data to be amenable to various types of statistical tests?

Sample Sizes Needed for Confidence Intervals around DIFFERENCES between Means

Most of the discussions just above relates primarily to confidence intervals about a summary statistic such as a mean. Again, for emphasis, it is not good to compare these types of confidence intervals to see if there is an overlap between the intervals in a pseudo-hypothesis test fashion. Although most beginning statistical textbooks cover confidence intervals about a single mean, some do not even mention a confidence interval about a difference between two means or medians.

Even if future data users don’t make the mistake of lining up confidence intervals about single means to see if they are different, but instead do it “the right way” (by

making the confidence interval about the difference between two means rather than about a single mean), then they are doing a hypothesis test identical to the normal kind, and in that scenario one needs to plan for adequate statistical power and all sample size calculators therefore need an input for beta.

To estimate sample sizes needed for confidence intervals between means, simply use two-sample hypothesis test (for the difference between two-means) sample size-calculators that have inputs for both alpha and beta. **Do not** use the simplistic single-sample confidence interval sample size calculators that have an input for alpha but not beta (for reasoning, see section below entitled “Parametric Confidence Intervals about a Single Mean”).

Using a hypothesis testing method to estimate needed sample sizes with a stated degree of power, when what is actually going to do is calculate confidence intervals rather than do a hypothesis test, is called the “power approach” by some. For reference and more details, the reader is again referred to a helpful internet resource (Di Stefano et al, 2005, op. cit.).].

The power approach is more appropriate for many NPS applications than the so-called contrasting “precision approach” (sic, what statisticians tend to mean when using the word precision is really about how wide a confidence interval is, with wider CIs equaling less precise estimates of CIs) that De Stefano discussed.

NPS managers typically want to be precautionary in managing rare and valued resources. To accomplish that, it is safer to estimate needed sample sizes based on limiting beta to small levels, even if one is supposedly just calculate confidence intervals on means. Again, soon or later, others will begin lining up confidence intervals about means and interpreting them in the wrong way. The only correct way to use confidence intervals in a context similar to a standard null hypothesis test is to calculate a confidence interval around **the difference between two means, rather than, (for example) a confidence interval about a single mean.**

Options for sample size calculators include 1) a two-sample sample-size calculators for differences between two means, and 2) a one sample sample-size calculator related to the difference between a mean and a single value (like a water quality standard).

Nonparametric Confidence Intervals about a Single Median

Nonparametric confidence interval (CI) estimates for the median are traditionally computed using the binomial distribution (Helsel, D.R. and R.M. Hirsch. 2002. [Statistical Methods in Water Resources](#). US Geological Survey Techniques of Water Resources Investigations).

Talk to your statistician before calculating a nonparametric (NP) CI using very low sample sizes. As in the case for parametric CIs, one issue is whether or not the full range of conditions of the [Target Population](#) was included in a relatively low number of samples. In other words, very low sample sizes often increase the probability that the CI will not be especially representative of the target population.

CIs should usually not be calculated if the sample size is less than 6-8. Also with such small sample sizes, the confidence interval is likely to be unacceptably wide for

many project goals and/or overlap impossible values (like zero, see further discussion below).

A handy rough first estimate of nonparametric 95 or 99% confidence intervals about a median for sample sizes of 6-120 can be approximated with the [UNB tables](#). If one has fewer than 6 observations, trying to do advanced inferential statistics (including CIs) on those few numbers is often unjustified and akin to “much ado about nothing.” In other words, one cannot create substantial new information content where very little information content exists. Trying to read too much into a CI from extremely small sample size (or pretending that a standard error from sample size of 2 or 3 means something), can be a sign that the author either understands little about statistics or is trying too hard to stretch the meaning of anecdotal results.

Once one has moved into the final stages of planning monitoring, if enough representative and credible preliminary data is eventually available, a statistician can help better estimate required sample sizes through bootstrapping simulations. The use of bootstrapping to estimate variance and “confidence intervals that are free of normal distribution assumptions” is discussed on a generic (terrestrial focus) FS document reproduced on a NPS VS website ([Statistical Techniques for Sampling and Monitoring Natural Resources](#) by Schreuder et al. 2004).

However, keep in mind that “There is considerable controversy concerning the use of bootstrap confidence intervals...Jackknifing and bootstrapping are no remedy for an inadequate sample size. For nonparametric resampling methods, the sample distribution must be reasonably close in some sense to the population distribution to obtain accurate inferences.” (W. Sarles 1995 SAS [bootstrap confidence intervals](#)).

Some say that bootstrapping techniques are not recommended unless sample size is over 40 (Elzinga et al. 1998. [Measuring and Monitoring Plant Populations](#)). Others would say that the caution regarding over 40 is overly cautious and not widely held. The key issue is that a bootstrapped CI based on a sample size of 15 is still a result based only on a sample size of 15. No matter what method for computing the interval is done, it will be worse than one based on $n=20$. If those 15 don't cover the full range of conditions, the CI will be too narrow. However, bootstrapping is generally recognized as a better way to compute a CI for small samples than standard parametric formulae. So it's the best of a bad situation, but is not necessarily to be totally avoided (Dennis Helsel, USGS, Personal Communication, 2006).

Sample Sizes and Statistics for Taxonomic Richness

This is a complicated subject, see your statistician. For a brief introduction, see Oregon State University Statistics [Urquhart summaries](#).

Sample Sizes Needed for Trend Analyses

In long-term monitoring, one is typically not just interested in documenting status, but also trends. The first of five generic VS **Goals of Vital Signs Monitoring**, and one embraced by most NPS VS monitoring networks is to:

[Determine the status and trends in selected indicators.](#)

How both status and trends will be determined should be summarized in the data analyses SOP. General discussions of statistical options for trend analyses, and a summary of the frequent need to consider diel, seasonal, flow, or phonological co-factors are included herein in the [section on a Data Analysis SOP](#). Here, the focus is narrower, on sample sizes needed for trend analyses.

Estimating sample sizes needed for trends can be tricky, and hints of some kind of trend can almost always be found in long term monitoring.

However, this should not prevent one from looking at the data from different angles, including relatively simple ones. In other words, even if one is going to look at trends using relatively sophisticated analyses, it is often good to first (and/or also) look at potential trends in relatively simply ways. For example, one can use [sample size calculations for paired t-tests](#) between various logical time periods (especially before and after some event that might logically cause a step-trend) as a part of basic functional data analyses. This would logically be part of an initial [exploratory data analyses \(EDA\)](#) step, or simply part of a data summarization step. Thus, one logical first step in looking at sample sizes needed to trend analysis would be to calculate sample sizes needed to detect a difference between two means using paired sampling. Then keep those values for required sample sizes in mind as a common sense check when developing needed sample sizes using more sophisticated trends analyses such as those described below. If the result of a sophisticated analysis is a much smaller required sample size, double check to see if an input variable was formatted wrong or an assumption was broken in the more sophisticated analysis.

In a paper that clearly demonstrates the importance of long-term monitoring data to understand complex ecosystem dynamics and illustrates use of data from a variety of sources and makes extensive use of conceptual models to express hypotheses on ecosystem processes and dynamics, Sinclair et al. (2007) noted that very slow changes are often not apparent to those experiencing the trend. In fact, they sometimes become apparent only after several decades. On the other hand, even slow changes can result in irreversible changes into a new state in which the system can remain for long periods, perhaps until a new disturbance shifts it to yet a different state or back to the original state [Sinclair, A.R.E., Mduma, S.A.R., Hopcraft, G.C., Fryxell, J.M., Hilborn, R., Thirgood, S. 2007. [Long-term ecosystem dynamics in the Serengeti: lessons for conservation](#). Conservation Biology 21(3): 580-590].

Again, even though statistics used for trends are often nonparametric tests (like the seasonal Kendall Test popular in water quality analyses within the USGS and other agencies), required sample sizes are often first approximated with parametric calculators.

Monitoring networks should consider establishing minimum sample sizes and/or minimum periods of time monitored before a trend can be called. Such criteria could be part of criteria used before the network decides to call a trend.

For example, one Australian state will not call a trend unless 1) the sample size was adequate (according to the first equation described below, for [two-sided](#) trend applications, and 2) the result of a Kendall test for trends indicates a trend. These issues and others (autocorrelation, etc, are discussed in a plain-language Internet document (Western Australia Water and Rivers Commission. 2004. [Statewide Assessment of River Water Quality Methods](#)).

Both two-sided (trends either way) and one-sided (trend in one direction only) sample size calculators are available from EPA (beta version) [Sample Size and Sample](#)

[Frequency Estimator](#). Parks are typically interested in trends whether the trend is going up or down. Both one sided and two-sided approaches are summarized as follows:

Two-Sided Trend Applications: The equation used by EPA and others to estimate adequacy of sample sizes for trends is $n = 12 * (\text{sample variance of the de-trended series}) * [t_{a2(n-2)} + t_{b(n-2)}]^2 / \text{trend magnitude}^2$, where $t_{a2(n-2)}$ is the [two-tailed](#) critical value for the t-distribution for sample size n-2 and where $t_{b(n-2)}$ is the one-tailed (upper) critical t-value for sample size n-2 using alpha of 0.05 and beta of 0.1. In this equation, t_a refers to the critical t-value corresponding to the Type I errors, corresponding with alpha, and t_b refers to the critical t-value corresponding to the Type II errors, corresponding with beta.

De-trending techniques in Excel are explained in the discussion of [Time Series data analysis](#).

One-Sided Trend Applications: The [one-sided](#) equation (input variables bolded) is very similar but uses a one-tailed t value: $n = 12 * (\text{variance estimate}) * [t_{a,v} + t_{b(1),v}]^2$, all divided by the trend magnitude squared, where $t_{a,v}$ is the one-tailed (upper) critical t-value for sample size n-2 and where $t_{b(1),v}$ is the one-tailed upper t critical value for sample size $v = n-2$. As can be seen in the equations, in the [one-tailed](#) case (t alpha) has a right-tail area of alpha. In the [two-sided](#) test, t alpha has a right-tail area of alpha/2. Regardless of whether the one or two sided choice is chosen, in the EPA calculator (op.cit.), **significance** (example alpha = 0.05) and **power/detection probability** (example 0.9) may be changed in the fields just below the equation, which automatically changes both the alpha and beta terms accordingly. The one-tailed equations are of special interest to EPA for regulatory questions, such as “Did best management practice implementation improve historically poor water quality in a watershed by some given percentage?” or “Did a new industrial discharge result in declining water quality?”. **If a suspicious hint of a trend appears to be in one direction only (the trend line is consistently going mostly in one direction)**, for the one-sided trend equation and calculator, see EPA beta test (in the EPA Version 0.7.2.2) [sample size calculator](#) for trends (choose trends and then choose linear trends).

Complex Ways to Estimate Variance for Trends Analyses:

The goal is **not** to get an accurate one time (status) estimate of the variance based on unequal probability of selection designs (such as [GRTS](#)). Instead, the entire purpose is to get variance component estimates that can be used to evaluate alternative design options for surveys over time. In this case, the local neighborhood variance estimate is not used as discussed above for status, but instead restricted maximum likelihood variance estimates are made. As was the case for the local variance estimate for status, maximum likelihood variance estimates are complex enough that an applied statistician may be needed to assist the monitoring networks to make sure the calculations are done in optimal ways.

For example, in a study of salmon habitat trends, Larson et al 2004 separated individual variance components to see how survey design detail changes (and various combinations of revisit schedules vs. sample sizes and other details) would impact power to detect trends over time. Variance contributors were divided into 4 categories: 1) Residual Variance (Which includes lack of perfect measurement [precision](#), changes caused by different observers, and very short term variation such as [diel](#) variation), 2) variation between sites, 3) variation between years, and 4) and interaction variation. Larson et al explained how one can vary survey design details and see how the changes impact variance (based on a root sum of squares method that seems to result in a type of pooled sample variance). The variance squares to be added included the 4 categories of variance listed above, with some details and various weighting schemes depending on monitoring design details. So the goal in this case was to define optimal monitoring design details. Another conclusion of interest was that “individual variance estimates for each survey did not differ in any substantial way from the grand estimation” Habitat variables tend to be less variable than water column water quality variables, but Larsen et al. 2004 still concluded that 30-50 samples per year would be ideal in detecting long term trends of these variables (Larsen, D.P., P.R. Kaufman, T.M. Kincaid, and N.S. Urquhart, 2004. Detecting Persistent Change in the Habitat of Salmon Bearing Streams in the Pacific Northwest. Canadian Journal of Fish and Aquatic Sciences 61:283-291, [Abstract](#)).

A related but older paper is Larsen, D. P., T. K. Kincaid, S. E. Jacobs and N. S. Urquhart 2001. [Designs for evaluating local and regional scale trends](#). Bioscience 51:1069-1078). The paper by Larsen et al. 2001 considers two components of variation: within sample interval (e.g., year) and across years but does not specifically try to separate out [diel](#) variation (one way to stratify by time of day to get variability down) or contributions to total variation from lack of perfect measurement [precision](#). The applications of these to local data can be complex.

Other resources related to varying monitoring and revisit details in ways that influences sample size and power calculations) can be found: 1) In additional references in [Part B](#), 2) in the presentation by Urquhart (S. Urquhart, 2006, [Sampling Design Considerations](#) at the San Diego National NPS Vital Signs Meeting in 2006, which discussed not only variance aspects but gives examples of statistical power for various rotating panel revisit options. Another older but often quoted reference is Larsen, D. P., N. S. Urquhart and D. Kugler. 1995. Regional scale trend monitoring of indicators of trophic condition of lakes. Water Resources Bulletin 31:117 – 140.

A systematic sample is a type of cluster sample, because once you pick the first point, all the other points are determined. The problem with compact cluster samples is that the variance one gets from such samples underestimates the variance of the underlying [Target Population](#), often seriously. We do not like to use the systematic random sampling variance with systematic samples because it overestimates the variance (Paul Geissler, Patuxent USGS BRD, Personal Communication, 2006, see [Part B](#)).

Another way to approach trend analysis is to estimate variance variances when testing for change between two time periods, for example a difference in two means. Depending on project specifics one could use a simple variance estimate of total variance for each of the time periods if inclusion probabilities were equal. However, for the unequal probability options (such as [GRTS](#)), one would instead tend to use either:

1. For means only (not CDFs) an alternative variance based on Horvitz-Thompson variance formulas (requires approximation to joint inclusion probabilities).
2. A more standard “local variance estimator” of variance.

Downloadable tools for both of these options are available from the EPA [spsurvey](#) (op.cit, citation above).

Rethink Detectable Difference Goals for Trends

Once one has looked at possible trends from some of the angles discussed just above, it is often good to rethink how big of trend magnitude (as a minimum detectable difference) the proposed monitoring needs to be able to detect.

EMAP tries to detect a 20 % minimum detectable difference in means over 10 years. In other words, they want to be able to detect a 2% a year change over 10 years. In another EPA example, one criterion used for picking an indicator was whether or not it could detect a 20% change in ecological condition over a 10-year period with 90% confidence (Kurtz et al. 2001, op. cit., citation above in objectives section).

Again for emphasis, run tentative [MDD](#) goals by park management to see if they are acceptable related to resource management needs. For highly valued or rare species not characterized by very high natural variability, superintendents have sometimes been reluctant to accept being able to detect changes of 50% in one year. In some cases, they have understandably been reluctant to be on record as being willing to accept 50% losses without even knowing the loss had happened.

Keep applying common sense tests to all decisions related to trends. In local or regional areas where there are wet years for several years and then several years of dry years, baseline monitoring may have to be done a long time to establish what is normal. Climate and other changes have a way of redefining normal. This fact should be taken into account when one is trying to detect trends; even if one is using complex methods such as [multivariate control charts](#) (see [What are Multivariate Control Charts?](#)).

How long does the time frame logically need to be to define conditions outside of normal conditions at relatively pristine sites?

Cautionary note on control charts in general: In laboratory measurements of chemicals, quality “control charts” have commonly been used to warn operators that the measurement process has become “out of control” (probably unreliable and in need of new calibration). Baseline data to generate such control limits are usually based on long term (read high sample size) lab performance for: 1) repeat measurements of the same thing ([precision](#)) and 2) especially for [bias](#)/systematic error (% recovery, only truly “accuracy” for longer-term estimates from sample sizes exceeding 25). In these [QC](#) applications, distributions are sometimes (relatively) normal, so empirical rule-based (multiples of the standard deviation) control limits have often been used. Even for these applications, some have pointed out that nonparametric alternatives are better, since environmental data is almost never normally distributed. In fact, USGS has at times (notably from 1999 to 2005) used F-pseudosigma instead of a standard deviation to estimate both control and detection limits (see USGS discussion of [Long Term Detection Levels](#)). Using control charts relative to multiple measures of variables in the

outdoor environment is a whole different scenario and can often be even more problematic than typical lab [QC](#) uses, especially if non-normal distributions and/or or small sample sizes (not long enough to establish a valid baseline) are not properly taken into account. The concern is even larger (red flags should go up!) if multiples of the standard deviation (rather than proper standard error-based t-distribution confidence intervals or various nonparametric alternatives) are utilized.

Additional discussions of the limitations and common misuse of the empirical-rule-based intervals (multiples of the standard deviation) and documentation that environmental data is seldom normal are in Helsel and Hirsch's statistical text book (Helsel, D.R. and R.M. Hirsch. 2002. [Statistical Methods in Water Resources](#). US Geological Survey Techniques of Water Resources Investigations).

For many taxa with large fluctuations in pristine environments, a 10% or even a 20% local change in 10 years would be impossible or costly to detect, and one would not usually go to the trouble and expense if the taxa were not endangered or threatened. Local changes of 50% or even more in 10 years even in pristine sites are perfectly natural for some highly variable species or groups. For extremely variable groups (bacteria, zooplankton, etc.) changes much higher than 50% are normal. Other things being equal, long term monitoring groups rightly tend to avoid monitoring extremely variable parameters. Although the endangered species criteria above apply globally, they may be of some interest for rough comparison with goals of how big of a change one would like to be able to detect locally, especially when one is dealing with relatively rare or threatened resource.

At the other end of the spectrum, is interesting to compare the kinds of trends that need to be detected for vertebrate endangered species. A new tool is available for trends. An equivalence test has recently been developed for demonstrating the absence of a trend. Sample sizes can sometimes be insufficient relative to the residual variation (and perhaps also autocorrelation) to call a trend. Results from equivalence tests depend critically on the magnitude of the equivalence interval. In one example, a half-life or doubling time of 20 years for population size was discussed for long-lived and relatively stable species. In an example discussed as more appropriate for shorter life-spans and more variable species, a less conservative equivalence region corresponded to a halving or doubling time of 10 years. In an example that used amphibians species on a global (not park) scale, simplifying the definitions of The World Conservation Union slightly, a decline in numbers of >50% in 10 years was said to define an "endangered" species and a decline of 30% in 10 years defined a "vulnerable" species (P. M. Dixon and J.H. K. Pechmann. 2005. [A statistical test to show negligible trend](#). Ecology, 86(7), pp. 1751–1756). However, in the NPS, we would ordinarily not use equivalence testing, but rather the more conservative and precautionary [inequivalence testing](#).

Calculating needed sample sizes for trend detection can be complex and whether or not a trend is detected can depend on numerous indirect details. For example, trends detected using a low laboratory detection limit (usually a [MDL](#)) and the (now-discredited) practice of censoring [nondetect](#) data to one half the MDL; were often not

detected when a higher MDL limit was adopted. Changing from monthly to quarterly sampling frequency resulted in fewer trends being detected. The details of how one adjusts data or weights data versus flow can also influence the results (as well as explain some patterns) and need to be considered (B. Stansfield, 2001. [Effects of sampling frequency and laboratory detection limits on the determination of time series water quality trends](#). New Zealand Journal of Marine and Freshwater Research, Vol. 35: 1071-1075).

The USGS has published a whole series of documents that drive home the point that many water column parameters are driven strongly by flow intensities. Many USGS publications therefore utilize [flow-weighting](#). High flow often changes the magnitudes of many parameters, and normalizing or weighting data by flow, load, or yield can be helpful in comparing sites or explaining variability patterns.

13) When In Doubt, Throw It Out:

This theme is so important that it is brought up multiple times herein. See additional related discussions in Section III (above) on documenting how measures and vital signs were picked. On the scale of multiple measurements, retain only measures that have acceptable minimum detectable differences over stated periods of time.

On the scale of each individual measurement, consider using only measures having acceptable levels of measurement [precision](#), acceptably low detection limits (or good [sensitivity](#) as AMS), and acceptably low measurement [bias](#).

Once calculations have been done on required sample sizes, this topic should be revisited. Measures or strata with excess variability will often prevent detecting a trend or a difference of even a large magnitude with existing budgets. Upon discovering that initial plans for the monitoring design (including what/where/how often to monitor) will not result in being able to detect a difference of concern, adjustments usually must be made.

Often monitoring planners throw out measures and strata that are obviously too variable to ever detect an effect size of concern with available budgets. If there is a strong desire to keep the vital sign or measure, try stratifying to get variability and sample sizes down to reasonable levels. If measurement uncertainty is excessive due to poor measurement [precision](#) or excess measurement [bias](#), adjust the field or lab methods to get the uncertainty down to acceptable levels, or choose a surrogate/alternative measure that can be quantified with less measurement-level uncertainty.

When the summary statistics of concern are **means**, consider throwing out analytes or measures where the variability in the sites sampled, is so high even in pristine sites (and even using strata or response design details that reduce variability the most) that one would never find a trend or difference of biological concern given funding limitations.

For the calculation of **proportions or percentages**, the thought process is a bit different than for means: If one cannot get a sample size of 25-50 in a logically defensible single sample (a sample that has a good chance of covering the full range of conditions of the named [Target Population](#), defined in detail in time and space), consider throwing out the variable or making other changes in approach. For single proportions, see EPA discussion of ““[Why a sample size of 50.](#)”

Don't Just Report an Unsatisfactory Result, Change Something:

If one finds out there is only an 11% chance in detecting a very large (and presumably ecologically devastating) change over 50 years, then “change something!”

This usually means either throw out the measure or indicator or find some way to get variability down.

If one cannot get variability down (usually by restricting sampling and inference to smaller areas in time and space), perhaps just abandon the measure. After all, most networks have far more candidate measures than funding to support them all. Don't just continue as planned because some assembled group thought that monitoring variable X might be a good idea. Most likely, they were unaware of the problem of detectable differences, or they might not have recommended variable X.

However, a counter argument that can be relevant at times is that there may be cases where one considers statistical issues and logically decides to stick with a measure rather than throwing it out, simply because other factors are concluded to be more important. No amount of statistical sophistication can make up for missing the most important variables (or bad data, or a bad monitoring design, or not meeting required assumptions, for that matter).

Thus, statistical considerations should not automatically override all other lines of evidence related to drivers impacting biological or ecological factors (think twice before throwing out oxygen or pH).

In one relevant example, the Northern Colorado Plateau Network decided to still measure TP even though [MDDs](#) were less than ideal, because (Dave Thoma, NPS, Personal Communication, 2007):

- 1) TP is part of a free nutrient analysis suite done for a cooperative agreement with the state.
- 2) At a screening level the Network would be able to see if something drastic was occurring in sites where I don't have sufficient historic data to do power analysis.
- 3) The Criteria for listing on the 303(d) list are not based on statistics. Listing is based on a fixed number of exceedances in a 12 month period.
- 4) The water quality standard of concern and 303 listing is for TP.
- 5) Some the variability of TP is due to ambient levels being near [MDL](#) detection limits (unavoidable, it is what it is).
- 6) In spite of MDD issues, TP is probably the best “total phosphorus load” parameter.

The above is an example of a good justification of why a measure was chosen in spite of being more-than-optimally variable even in pristine environments. In a related note, phosphorus parameters tend to be highly dependent on flow, so there might be a better chance to detect trends if the P data collected is [flow-weighted](#) prior to trend analyses.

14) Optimize Monitoring Plan Details for Affordability and Logic

Monitoring design optimization steps not only include throwing out measures with excess variability, but also restricting [Target Populations](#) in time and space, not doing monitoring already done by others, and considering other steps that could be done to optimize monitoring. For example, if detection probabilities are still too low after the steps above are completed, consider the following:

Often the choice is between: A). monitoring many sites and measures very infrequently and poorly, or B). monitoring fewer sites and measures more rigorously and/or more often. Choice A has too often resulted in data that can be used for very little (if anything) related to management decisions or regulatory purposes. Choice B is sometimes a better option that produces at least some useful information. During plan optimization steps, reconsider the overall affordability and logic of sample sizes, sample placement, sample replication (how many samples at each site, where to sample, how often to sample, when to composite (or not), statistical significance, and statistical power.

The goal is to come up with a combination that will produce acceptable detection probabilities and will produce information (not just data) useful to park managers for resource management decisions. The more (and the earlier) the network quantitative ecologists and statistical consultants can help with these steps, the better.

15) Draft Initial Sample Sizes and Optimized Monitoring Design

Also assemble the best available estimates for input variables (standard deviations, alpha, beta, see list above) to take to the applied statistician (next step).

16) Finalize Sample Sizes and Design with an Applied Environmental Statistician

Once network quantitative ecologists and small groups of specialists that are finalizing protocols and SOPs have completed the steps above, they should strongly consider consulting with an applied statistician, taking that expert the correct information and input variables (above), refined questions (detailed in time and space), and refined [Target Population](#). The generic basic design developed for the earlier Phase II report may have been envisioning larger sample sizes and the assumptions may have changed. If the sample sizes have been cut and other changes have been made in design optimization steps taken when developing QA/QC and data analysis SOPs, the revised plan needs to be checked again by a statistician. Typically the first version of chapter 4 (Monitoring Design) of the central monitoring plan is drafted a year or more before the SOPs are finalized. In the next year there have often been disconnects and assumption changes between earlier statistical advice and later changes at the protocol and SOP detail development stage.

Distributions are typically not normal, samples are often not large, and various assumptions may not be defensible. Standard power and sample size analyses may get one in the ballpark, which will usually be adequate given that preliminary data that one bases the calculations on are often not optimal. However, if preliminary data does cover the full range of conditions, some analyses are too complicated to rely solely on plug-in power calculations. Sometimes, multiple hypotheses need to be considered simultaneously. This requires more complex methods, such as Monte Carlo simulation-

based approach to determining sample size and power (see P. Lukacs 2005 [Beyond Simple Power Analyses](#) from the NPS VS Austin, TX Meeting).

Remember however, that you need to take meaningful data to the statistician. The initial data available before the start of simulations must have sample sizes large enough (look at the data closer if the sample size is less than 30-200) to be optimally useful in simulations. The initial data must also be relevant and representative of the full range of time and space conditions of the [Target Population](#). That last caution also applies not only to simulations and other complex calculations but also to simple-algebra Zar sample size calculators.

After completing consultations and final checks with an applied environmental statistician, finalize the following in the protocol narrative and SOPs: 1) sample sizes, 2) minimum detectable differences (or alternative target effect sizes), and 3) sample placement in time and space detail. After monitoring designs are modified and finalized in [optimization](#) steps Chapter 4 (monitoring design) of the central monitoring plan for each Vital Signs Monitoring Network will also need to be modified to reflect the final design.

The sample size needed to determine a desired minimum detectable difference (and how it was determined) relates to many other issues. Therefore, we suggest networks not only document how [MDDs](#) and target [thresholds](#) were determined in the data analysis SOP, but also include brief recaps or “point to” links in other related sections, such as the discussions of [representativeness](#) and [completeness](#) in the QA/QC SOP, and the sampling design discussions (Chapter 4 in the central monitoring plan).

17) Estimate the % of Samples That Will Fail

It would be rare for 100% of planned outdoor environmental samples to produce useful data. Seldom are all planned samples successfully obtained and also pass all [QC](#) data acceptance criteria. Samples or samplers may be lost in the field, lab or field complications may interfere, and samples can get lost or be spoiled while being shipped to the lab. Weather events may interfere with sampling or analyses (when shipping samples to coastal areas, hurricanes have delayed analyses), staff or equipment failures can be a problem, or delays may cause maximum holding times to be exceeded. A new technician might also use the wrong type of container or otherwise contaminate samples. Therefore, before required sample sizes are finalized, one first needs an estimate of the % of planned samples that may fail. If no other good rationale can be developed, planners sometimes pick a number like 10 or 15% to start with and adjust it as experience is gained.

18) Increase the Planned Sample Sizes Accordingly

Next, adjust required sample sizes upward to correct for the % expected to fail. For example, if 15% of the samples are expected to fail, multiply the required sample sizes developed in 16 by 115%, and edit the plan, protocols, and SOPs accordingly.

19) Include Completeness Goals in a Table in the QA/QC SOP

For each parameter to be measured, include a completeness goal in the SOP. If 15% of the samples are expected to fail, put 85% in the [table](#) as a completeness goal.

End of completeness, sample size, and statistics vs. desired conditions outline and chapter.

VII. Data Comparability (Internal/NPS and External/Other Regional Data)

We are now moving from QA topics to [QC](#) topics. Comparability is usually considered a QC basic, albeit one assured qualitatively. More statistical tools are now being developed, and in future years the comparability may be assured more quantitatively.

For Internal Data Comparability: What will be done to maximize temporal and methodological consistency in NPS data? Control typically involves limiting changes in internal NPS methods or timing of sampling to help insure our own newer data is comparable to our older data. However, due to advancing technology and other factors, changes in both methods and personnel are inevitable. When such changes occur, any resultant measurement [bias](#) from the change should be documented in the [Cumulative Measurement Bias SOP](#). The question then becomes: Is our internal NPS data from both old and newer measuring systems comparable enough that the different sets of data could be combined for purposes of determining trends or making management or regulatory decisions? If not comparable enough for that purpose, newer data and older data will often need to be normalized to data as of one (baseline, often starting) date. For more information, (see [Include a Cumulative Bias SOP](#) section below)

For External Data Comparability: What will be done to achieve comparability with other regional data sets such as [ECDMS](#) and will labs approved by other federal agencies. For example, will FWS approved labs ([Analytical Control Facility](#)) be used? Will an effort be made to standardize with USGS, State, NOAA, or EPA CERCLA site methods and/or labs? Are exactly the same parameters and fractions (total or dissolved, for example) being measured in identical media, and are Measurement Quality Objectives for [precision](#), [bias](#), and [sensitivity](#) (sensitivity is usually expressed as a [MDL](#) low-level detection limit or as [AMS](#)) similar enough to ensure data comparability? What will be done to insure our NPS data are comparable enough to the data from other state and federal agencies that need to be convinced our data is credible and comparable, given our purposes for monitoring? Is our NPS data comparable enough to other important outside data sets that the two sets of data could be combined for purposes of determining trends or making management or regulatory decisions?

Has the chemical lab proposed for use 1) passed federal round robin blind sample checks (see [FWS example](#) at), or 2) performed acceptably in other federal round robin blind checks (see [USGS example](#)), or 3) been approved to work for the parameter of interest and media of interest by the Federal [National Environmental Laboratory Accreditation Conference](#) (NELAC)?

For sediment or fish tissue monitoring, a good way to ensure data comparability with the large FWS nationwide data base (on metals, pesticides, herbicides, PAHs and other oil-related compounds, dioxins, PCBs, and other toxic chemicals in fish, wildlife, sediments, and soils) would be to use the same [FWS-approved contract labs](#), or at least

ask for the default (or at minimum, at least as stringent) [FWS QA/QC measurement quality objectives](#) for [precision](#), [bias](#), and [sensitivity](#) as detection limits.

For water column parameters including the nutrients, the [QC performance of labs that participate in the USGS SRS round-robin comparison](#) can be checked to see if the performance meets monitoring network target measurement quality objectives.

Among labs that have passed federal QA/QC checks, it is often optimal to choose labs that have produced a large amount of data for other federal programs to gain maximum data comparability. For example, if the network is producing water column data, the USGS NWQL lab or [USGS round robin test labs](#) might be among the optimal choices. If the network is producing data on benthic macroinvertebrates, the USGS labs used by NAWQA or the [Utah State Bug Lab](#) (especially for western or BLM data) labs might be good choices. A key question is: which labs have produced the most data that will be compared to new data being generated by the NPS network?

A final check should be made to make sure both the lab and the field method SOPs attached to the protocol are detailed enough to allow for reproducibility of exactly the same methods by third parties. Are they also detailed enough to allow judgments about the comparability of the data with the data of other agencies? Perfectly comparable data can be merged and analyzed together without introducing problems.

These issues are just as important for biological monitoring as for water chemistry monitoring. Interagency efforts are now being made to come up with acceptance criteria to determine data comparability.

A recent document explains many bioassessment **data comparability** issues for **large river** monitoring but the methods for comparability analyses considering Measurement Quality Objectives ([MQOs](#), including [sensitivity](#), [precision](#), [bias](#), and [precision](#)) seem broadly **applicable to smaller** (wadeable) rivers as well (Flotemersch, J. E., J. B. Stribling, and M. J. Paul. 2006. [Concepts and Approaches for the Bioassessment of Non-wadeable Streams and Rivers](#). EPA 600-R-06-127). Therein, accuracy is (appropriately) identified as a different concept than bias. Not mentioned, however, is that the type of accuracy explained (the proportion of times a scoring system accurately classified a site) typically requires a sample size of 25-50 for a good (confidence interval reasonably small around the proportion) estimate of the proportion. Another concept discussed of interest in the NPS is that “programs needing only to separate extremely disturbed from minimally disturbed sites will require less precision than programs designed to detect small departures in ecosystem condition.”

Comparability in Agreement or Pass/Fail Scores

In a topic somewhat related to the paragraph just above, as of 2006, there are no universally accepted ways to assess “agreement” (a different topic than correlation) in ratings or scores for biotic “condition,” some of which have only two possible ratings (pass or fail). Highly correlated scores (such as index of biotic integrity scores) indicate high “association” but do not guarantee a strong strength of “agreement.” For example, results from one state sampling protocol may rate stream condition consistently one level higher or lower than that of another state or federal program. In this case, the strength of agreement is not strong, although the correlation/ association may be very strong.

Accordingly there has been much recent interest in various options for measuring strength of agreement.

It is easier to make comparisons with scores or ratings when there are only two choices rather than with many different categories. These matters are the subject of much current interest. Again, it is much easier to make comparisons with two variables or two scores than with many. Networks should probably consider looking at such issues from different angles, including some relatively simple and intuitive ones.

It helps if dichotomous decisions can be made. Even in a relatively complex rating system (very poor, poor, fair, and good) the key goal might be maintaining “good” condition. Simple calculations could be made relative to the % agreement where both methods resulted in “good” scores, the [relative percent difference](#) between two scores, or the % bias comparing results of one method to another.

As another example, what % meets quality standards (pass/fail)? With enough data, networks can usually quantify the proportion of river miles that either passes or fails water quality standards. These values can sometimes be compared with other category dividers (say between fair and poor) in more elaborate systems.

. This approach would be consistent with a reality-check step of looking at the issues from different angles, including some relatively simple and intuitive ones. In other words, use intuitive and simple lines of evidence in addition to exotic coefficients (such as [kappa](#)) when possible

When there are multiple ratings (very poor, poor, fair, and good), things get more complex. Some have suggested using [kappa](#) or weighted kappa to look at agreement of IBI scores or to evaluate agreement in ratings of stream condition. For example, the EPA summary of the [Mid-Atlantic Integrated Assessment](#) Maryland case study took this approach.

The free McBride Cohen’s [kappa calculator](#) can be used as one way to look at agreement of dichotomous data. McBride also has a [Lin’s concordance calculator](#) on the net to calculate [Lin’s "concordance correlation coefficient."](#) This statistic appears “to avoid *all* of the shortcomings” associated with the usual procedures (such as a Pearson correlation coefficient) and can be used as one way to look at strength of agreement of continuous data.

However, kappa and similar methods tend to be complex and they have their detractors. Detractors say kappa is over-used. They also point out assumption complications and that not everyone agrees on how high a kappa score has to be to reflect various gradations of agreement. So agreement coefficients are something to have your applied statistician approve before finalizing them for particular applications.

VIII. Measurement Sensitivity

When measuring water quality chemicals at very low levels, a system that can accurately measure and detect a very low concentration is more sensitive than one that can only detect the presence of the analyte at higher concentrations. The more sensitive the measuring system, the lower the low-level detection limits are.

For chemical lab analyses, low-level data quality indicator (DQI) sensitivity goals or requirements for sensitivity are usually expressed in two ways: 1) a semi-qualitative method detection limit ([MDL](#), usually expressed in metric units such as mg/L or mg/kg)

and 2) a quantitative detection limit ([ML](#), usually 3.18 times the MDL). Actual MDL levels that can be achieved by any one lab vary a bit over time. One criterion for choosing a lab should be whether or not the lab can consistently achieve MDLs below the minimum level needed for project objectives. As first mentioned in the section on picking methods and SOPs, there should be an emphasis on picking methods and labs that can consistently achieve a semi-quantitative MDL detection limit lower than the lowest water quality standard (including chronic standards) or other "safe levels" or other [threshold](#) benchmarks of concern (optimally at least 1.6 to 2 times lower).

How low the detection limits need to be also depends on the concentration typically found in the environment being sampled. For example, down-gradient from cities or intense agriculture, concentrations of nutrients like TN, TP, Nitrate, Phosphate, TDN, TDP, etc. may be high enough all the time that one need not have the lowest possible (and therefore sometimes more expensive) detection limits to get say 5 times below the lowest levels commonly found in the environment. In pristine oligotrophic lakes in the National Park Service high altitude parks, the levels in the environment, and therefore nutrient MDLs, may need to be much lower (see [Oregon State example low level MDLs for nutrients](#)).

After monitoring starts, MDLs (lab work) or AMS (field measures) or the other relevant forms of sensitivity should be checked every so often to make sure original project goals for measurement sensitivity are being met.

Toxic chemicals can be hazardous at very low levels, and there is typically a concern about whether or not they are present in parks, even at very low levels. Likewise, some pristine waters in the NPS have very low concentrations of nutrients, and the parks want to keep them that way. Both of these scenarios lend themselves to documenting and controlling measurement sensitivity with the lowest-practicable detection limits. Low level detection limits have been the most common way sensitivity has been handled in the past, and for water quality parameters sometimes present in very low amounts, they are still critical.

If we are always measuring in higher measurement ranges (well above the low-level quantitation detection limits) we still need to control measurement sensitivity, but not always low level sensitivity. For some parameters measured in the field (pH, temperature, conductivity, biological observations, physical habitat observations, etc.), one seldom (if ever) encounters extremely low levels.

In these mid (or quantitative) measurement ranges, the smaller the (true) change that a measuring system can accurately detect, the more sensitive the measuring system is. For these types of measurements, low level detection limits are less relevant and/or less helpful. In these cases, we recommend alternative measurement sensitivity ([AMS or AMS+](#)) method to estimate measurement sensitivity can be used, as explained in more detail farther below, after the various low level detection limit sections.

As discussed in more detail below in a section on [AMS for biological or physical habitat measures](#), such measures are amenable to controlling sensitivity at the QC level as [AMS](#). Sensitivity as [MDLs](#) would be less optimal for most biological or physical measures, since very-low-signal-strength scenarios would tend to the exception rather than the rule

[Minimum detectable differences](#) are discussed elsewhere herein. That is because MDDs are about sensitivity at a higher level (the survey or overall monitoring design level) rather than the QC level being discussed in this section.

Low Level Detection Limits (MDLs and MLs)

In 2004, EPA clarified that (low-level) “detection” indicates the presence of a pollutant in a sample. Quantitation, on the other hand, indicates how much pollutant is in the sample (EPA 2007. [Procedures for Detection and Quantitation](#)).

Minimum Requirement: In the QA/QC SOP, list (a [table](#) is fine) pre-project targets (and how often they will be estimated once monitoring begins) for the following low-level detection limits:

1. A **semi-quantitative** method detection limit (MDL) and
2. A **quantitative** (minimum level of quantitation) detection limit (ML).

For NPS standardization, the MDL and ML are the suggested defaults. We suggest a minimum frequency for calculation of these values of at least once every six months or whenever there is a significant change in the measurement process.

Using the default suggestions given in many EPA methods would be acceptable. For example, many EPA methods suggest that MDLs should be determined every six months at a minimum. New MDLs should also be calculated more often than each six months when there is a significant change in the measurement process, such as: 1) a change in a measurement instrument’s responses to [blanks](#), 2) when [precision](#) or [bias](#) changes, 3) when a new matrix is encountered, or 4) when the lab believes (for whatever reason) that there may have been a change of low-level measurement [sensitivity](#)..

For example, when a new operator begins work or when there is a significant change in the measurement process (new method or new instrument), one should suspect that sensitivity may have changed and therefore recalculate MDLs. Some labs take the precaution to calculate MDLs more often than every six months or after significant changes, and that fine, a more precautionary approach to make sure the measurement process is remaining consistent (producing comparable data) and in control.

What does one put in the MDL and ML [tables in the QC SOP](#) in this situation? In many cases one can use another program’s defaults (for example, a state or EMAP’s MDLs). One might also just enter a code in the table that is explained in more detail at the end of the table. The code explanation might explain that as long as [precision MQOs](#) are met above twice the MDL and the field readings settle down to one value, these other-agency MDLs and MLs were considered sufficient. MDLs (or other forms of measurement sensitivity) need not always be the most stringent ones available, but they do need to be listed and meet project goals

The MDL and ML are to be calculated as follows:

MDL:

List the target standard EPA method detection limit (MDL), for each parameter to be measured, in a QA/QC SOP. Labs should be instructed to calculate the MDL as explained in Appendix B to 40 CFR Part 136—Definition and Procedure for the Determination of the Method Detection Limit—Revision 1.11). That same definition was reiterated, further explained, and defended by EPA in 2003: "[Technical Support Document for the Assessment of Detection and Quantitation Concepts](#)" (EPA 821-R-03-005, February, 2003), a very lengthy (200+ pages) updated recommendations and discussions of pros and cons of alternative detection limits past and present. The MDL definition has remained the same as the one in past years:

To determine the MDL, at least seven replicate samples with a concentration of the pollutant of interest near the estimated detection capabilities of the method are analyzed. The standard deviation among the replicate measurements is determined and multiplied by the upper ([one-sided](#)) critical t-value for n-1 degrees of freedom (in the case of 7 replicates, the multiplier is 3.143, which is the value for 6 degrees of freedom).

Although most labs and even some EPA staff and published EPA methods do not always use all the steps suggested by EPA to calculate a [MDL](#), most at least eventually use the central equation of Method Detection Limit (MDL) = t times S, where, t = the one-sided critical t-value for seven replicate (precision repeatability) samples. In this equation, for 7 replicate samples, t = 3.143, so MDL = 3.143 times the sample (n-1 version) standard deviation for the 7 replicate measurements of a [blank](#).

The same equation was used in the [APHA Standard Methods Book](#) definitions of a MDL, and by many states and others. The MDL is usually said to be the lowest concentration we really believe with 99% confidence is different than zero. The "different than zero" part calls for a one-sided statistical comparison, which is why we use the one-sided critical t-value. The same EPA equation is usually used to estimate estimated detection limits (EDLs), but with fewer steps than one uses in estimating a [MDL](#). Thus MDLs and EDLs are not usually the same value. Calculating EDLs is most often a preliminary step on the way to estimating MDLs. EDLs are often calculated with low-level standards or solutions rather than blanks.

When [blanks](#) never produce detectable signals, it is common to estimate MDLs in samples with the lowest possible levels where a signal can be detected ([EPA, 2003, Technical Support Document for the Assessment of Detection and Quantitation Concepts](#)."

To avoid confusion, alternative semi-quantitative detection limits (EDL, LOD, IDL, LLD, etc., see [Part B](#) for details) should ordinarily not be used instead of the standard EPA MDL. The exceptions include:

Some USGS labs have used the standard MDL, but the large NWQL USGS lab in Denver that produces a large amount of data for the USGS has typically used the Long Term MDL (LT-MDL) instead of a more standard MDL. For NWQL data, the LT-MDL should be listed along with how and how often it is calculated (The USGS NWQL used f-pseudostandard deviation rather than a standard deviation from 1999-2005, then went back to using the standard deviation, although some details in the calculation of

a LT-MDL were still different than a standard MDL). The LT-MDL requires a sample size of at least 24 rather than 7 (see [Long Term Detection Levels](#) and earlier more detailed discussion at [Open-File Report 99-193.pdf](#)). From 1999 to 2005, USGS substituted a “f-pseudosigma” instead of a normal sample standard deviation, both for LT-MDL calculations and for control charts. Most 1996 and later data from the NWQL is based on LT-MDL calculations that no longer use F-Pseudosigma (which was seen to reduce variability too much and to not be necessary since most of the data from repeat measures was close enough to normal to allow use a standard deviation, and since alternative ways were developed to handle outliers).

Note: The F-pseudosigma is a nonparametric statistic analogous to the standard deviation that is calculated by using the 25th and 75th percentiles in a data set. It is resistant to the effect of extreme outliers. Specifically, to get the F-pseudosigma, one subtracts the 25th percentile from the 75th percentile to obtain the inter-quartile range (IQR) magnitude and divides the result by 1.349 ([Long Term Detection Levels](#)).

If there is no good way to calculate or find a MDL but an estimated detection limit (EDL) can be found or logically calculated (sometimes the case for bacteria or chlorophyll) and calculated, the EDL can be defined and used. Some labs basically let electronic instruments define detection limits, since some instruments censor low level values in the noise range (below calibration curve limits). If this is the case, exactly how the semi-quantitative detection limit was determined, and why standard MDL calculations were not made, should both be documented in the sensitivity section of the QA/QC SOP.

[NEMI](#) sometimes gives the lower end of the calibrated range as a “range-derived” lower detection limit, and this or some other rough estimate of a MDL might be used in [QC tables](#) when bad [precision](#) when measuring close to a [MDL](#) does not require a more stringent calculation of a proper low-level MDL.

In cases where one is always two or more times above the MDL and never encounters really poor precision from measuring too close to a MDL, alternative measurement sensitivity (either [AMS or AMS+](#)) should be periodically calculated and reported.

How low should MDLs be? Labs picked should be able to achieve a semi-quantitative MDL detection limit lower than the lowest water quality standard (including chronic standards) or other chronic exposure [threshold](#) benchmark of concern [reference doses --- RfDs, No Effect Levels such as No Observable Adverse Effect Levels or No Effect Concentrations (NOAELs, NOECs), or any other “safe level” benchmarks]. Optimally MDLs should be at least 1.6 to 2 times lower than the lowest of any such benchmarks that can be found.

Minimum Level of Quantitation (ML)

Unless otherwise justified, define and calculate the minimum level of quantitation detection limit as 3.18 times the [MDL](#). The resulting value is the same as the (low level)

lower quantification limit (LQL, a STORET-specific term), but it is suggested that the ML terminology be used rather than LQL except when dealing with STORET. In previous versions of [Part B](#), this detection limit was referred to as a PQL, but 2003 EPA guidance documents make it clear that the ML (as 3.18 times the MDL) is a better default term for the quantitative limit [Technical Support Document for the Assessment of Detection and Quantitation Concepts](#).

The ML is typically very close or sometimes a bit lower than the also often-used limit of quantification (LOQ). The LOQ is normally 10 times the standard deviation (SD), the MDL is $3.134 \times \text{SD}$ and the ML is $3.18 \times \text{MDL}$. So for most practical purposes the LOQ and ML are basically the same ($3.134 \times 3.18 = 9.99 = \text{about } 10 = \text{value used for LOQ}$).

A more elaborate definition of the ML including rounding rules is also contained therein. The [2003 EPA document](#) also explains that the ML is "the lowest level at which the entire analytical system must give a recognizable signal and acceptable calibration point for the analyte. It is equivalent to the concentration of the lowest calibration standard, assuming that all method-specified sample weights, volumes, and cleanup procedures have been employed.

As explained in [EPA 2003](#), some of the criticisms of the MDL and ML relate to single lab vs. multi-lab comparisons. If a network needs to do so because it is dealing with multiple labs or cannot achieve quantitative detection limits at 3.18 times the MDL, the network could alternatively define a multi-lab PQL as 5 (rather than 3.18) as suggested in the [Standard Methods Book](#). Just make it clear that the PQL detection level is a multi-lab achievable rather than a single lab quantitative limit.

In the context of a single lab, there are disadvantages for using 5, and it should be a justified exception rather than a default choice. Using 5 would result in being able to report fewer low level values (see following section). Also, some EPA methods specify the use of 3.18. Top experts in the field now consider 3.18 to be sufficiently high to protect against false negatives, to be the value most commonly used, and to have other advantages over 5 (for details see [Part B](#), or D. Helsel. 2005. [Nondetects and Data Analysis: Statistics for Censored Environmental Data](#)).

To avoid confusion and for NPS standardization, use the ML rather than alternative quantitative detection limits phrases or acronyms. It is easy to become confused in the alphabet soup of the great many alternatives. Alternatives include practical quantitation limits (PQLs), minimum reporting levels (MRLs), reporting levels (RLs), limits of quantitation (LOQs), minimum quantitation limits (MQLs), sample quantitation limits (SQLs), Contract-Required Quantitation Limits (CRQLs), or an inter-laboratory quantitation estimate (IQE). These and many other variations are explained in [EPA 2003](#). Instead of using these terms, convert all such quantitative limits to MLs.

There are two exceptions where terms other than the ML can be used:

1. When dealing with STORET, the phrase "lower quantitation limit" (LQL) can be considered a synonym for a [ML](#).
2. If a USGS lab is used, the USGS alternative to the ML, the laboratory reporting level (LRL) may be used instead of the ML. The LRL is defined as two times the USGS LT-MDL (USGS. 1999. [OFR 99-193](#)). If the

network is going to use USGS LRLs, explain how and how often they will be calculated and reported.

No matter what quantitative detection limits are used, how they are calculated should be explained in the sensitivity/detection limit part of the QA/QC SOP. Once monitoring begins and new data is put in STORET, it should also be put in STORET metadata, as explained in the next section.

For toxic chemicals, it is particularly important that the lab can achieve ML quantitative detection limits that are below the benchmark, water quality standard or criteria, or other [threshold](#) levels known to be associated with harmful effects.

How Will Values below the MDL or ML be Reported and Analyzed?

Although this topic should be explained in the Data Analysis SOP attached to each protocol narrative, it is discussed here to keep it close to the discussion of detection limits (just above).

This topic could also be discussed in the [sensitivity](#)/detection limit section of the QA/QC SOP attached to each protocol.

Regardless of which SOP contains the discussion, there should be links back to the other for clarity, and the discussion should explain how data below any of the listed detection limits will be handled, not only for reporting into data bases, but also for data analyses. Along with how missing values will be handled (see [completeness](#)), how values below various detection limits will be handled should also be covered in the data analysis SOP.

One acceptable option (and one already adopted by some VS networks) is to state that the recommendations in the recent Helsel Book (D. Helsel. 2005. [Nondetects and Data Analysis](#)) will be used. Among other things, this book explains why one should not substitute one-half the detection limit for nondetects. For a brief Web available explanation, see "[Why substituting one-half for less-thans is a really bad idea.](#)"

NPS data needs to go into STORET, and Helsel ([Nondetects and Data Analysis](#)) considers the modernized STORET default recommendation for writing to a database to be fully acceptable. Therefore, we are adopting this as a default NPS recommendation. Modernized STORET and NPSTORET both suggest that we not report into a database any value higher than the MDL but lower than the [ML](#). Instead, the detection condition field is set to "Present, below Quantification Limit." With that detection condition, STORET automatically enters "*Present <QL" in the result field. A major advantage of this approach is that no "estimates" are treated as quantitative when they are not quantitative. NPSTORET is consistent.

A Kaplan-Meier calculation option for handling nondetects is available as an easy option to users of NPSTORET. Nondetect data (below the ML) are transformed (censored) to Kaplan-Meier transformed values first, and then certain summary statistics (mean, median, minimum, maximum, standard deviation, and percentiles, but not confidence intervals) are calculated in NPSTORET.

Those networks and parks that do not already have NPSTORET, and/or those who wish to do statistics other than those listed above can handle nondetects with the

Kaplan-Meier techniques with a free [KM MS Excel worksheet](#) available from [PracticalStats.com](#). Therein, Kaplan-Meier is a simple procedure that estimates the mean, standard deviation, median, 25th, and 75th percentiles, and the (t-interval) 95% [upper confidence limits \(UCL\)](#) for censored data. It is not a macro, just a simple worksheet, to minimize the possibility of viruses, etc. It has been checked and found clean by Norton Antivirus. It is limited at present to data at 100 different values - there may be multiple observations at each value, however. It 'flips' the data and performs all processing, so you simply put in the values and get back the results. If you find it useful, cite the source as PracticalStats.com when presenting or publishing results.

In (eventual) statistical analyses, values between the MDL and ML are best interpreted using either an interval-censored method (parametric) or a rank-based method (nonparametric). In the latter, all in-between values are represented as the same tied rank. The older recommendation of censoring to half the MDL is clearly no longer recommended.

For reasons consistent with both STORET and NPSTORET rationales, Helsel recommends that numerical values not be reported into data bases if the values are below the MDL or the ML, and that one should not report nondetects as half the detection limit. One should also not report nondetects as a negative (minus or -) sign followed by the actual MDL value, because someone invariably decides it really is a negative number ([Nondetects and Data Analysis](#)).

These recommendations are all followed in NPSTORET, which will not allow entry of values below the [ML](#). The MDL and ML limits are entered into NPSTORET, and by using STORET detection condition coding results, one can find out how many values were below the MDL or between the MDL and ML.

Values above the [ML](#) are classified in EPA's modernized STORET database with the detection condition of "Detected and Quantified" This is ideal, and according to EPA STORET Staff, this is optimally the only choice which permits reporting a single number.

Although not recommended in the Helsel book ([Nondetects and Data Analysis](#)), for the special case of NPS analyses of "precautionary principle" comparisons with standards or criteria, one might choose to censor all data below the ML to the exact value of the ML, but that is only a very special (worst-case, trying to be very precautionary and totally avoid false negatives) example of a data analysis strategy, and one would never substitute the value of the ML in a long term network storage data base field for measured concentrations.

Alternative Measurement Sensitivity (AMS) and AMS+

AMS and AMS+ are alternative ways to control and document measurement sensitivity at the [QC](#) level when low level detection limits (like MDLs) are not needed.

In the case of field measurements (pH, specific conductance, etc.) using electronic instruments, in the past if you asked someone:

"Since some differences between data points are small enough to be considered simply the result of measurement process noise (random up and down

measurement error) rather than a true difference, how small does a measured change have to be before you consider the difference to be real difference?"

The response might be to point to a manufacturer's "[resolution](#)" specification. For some biological or physical habitat observations, they might have even fewer ideas.

AMS is standardized solution and also a better term to use than the word resolution, a word that tends to mean many different things in water quality monitoring (each instrument manufacturer seems to estimate "resolution" differently). Also, for [QC](#) one needs to estimate actual performance in the field, rather than rely on a specification that relates to ideal performance in a lab. QC is performance-based, and if the measurements are done in the field, then the actual measurement performance should be estimated and controlled under field conditions as well.

In past laboratory analyses for low-level contaminants, sensitivity was often controlled by reporting [MDL](#) detection limits. In past laboratory analyses for higher level variables, no other control of sensitivity was commonly attempted. For many biological or habitat variables or estimates, measurement sensitivity was simply not controlled.

However, measurement sensitivity is an important [QC](#) topic and the need to control sensitivity does not disappear just because one never or seldom encounters very low-signal-strength or [nondetect](#) values. For vital signs monitoring in both the field and lab, we therefore suggest, that for any situation where low-level MDL detection limits are not optimal ways to control sensitivity, QC measurement sensitivity be estimated as AMS.

Just as measurement quality objectives for precision and bias should be listed in a [summary table in the QA/QC SOP](#) or [QAPP](#), so should measurement quality objectives be listed for AMS or AMS+ any time that MDL and ML detection limits are deemed inadequate to control sensitivity by themselves.

AMS calculations use a sample size of 7 and a confidence level is 99%, to be most functionally similar to [MDLs](#).

AMS calculations address the "[two-sided](#)" issue of how large of a difference between two measured values can be before we are 99% sure it is a true difference. For contrast, MDLs address the "[one-sided](#)" issue of how large a measured magnitude is before we can consider the value as a true "detection" (99% sure it is different than zero).

AMS calculations therefore use the "critical (two-sided) t-value" because we are not just interested in one-directional measurement sensitivity. MDL calculations use the "critical (one-sided) t-value."

Note: Those not familiar with critical t-values should read this paragraph. One-sided is a synonym for one-tailed. Two-sided is a synonym for two-tailed. In past versions of Part B lite, we have used the phrase "middle t-value." Although this phrase may help some visualize the distribution and choose the right one in the user-friendly [SurfStat Australia calculator](#), a more technically correct (and more common) way to say this is to not refer to the middle t-value at all but rather to call it the "two-sided critical t-value" or the "two-tailed critical t-value." Therefore we have discontinued using the phrase "middle t-value." To see how the "two-tailed critical t-values" change according to sample size and confidence levels, see the [t-distribution calculator](#). To calculate the proper value for sample

size 7 (DF = 6), type in 6 under the DF column, choose the two-tailed critical t-values choice (the middle of the distribution is red, choose the third distribution from the right), and then to get 99% confidence for two-sided cases, type in 0.99 under probability. Next, click on the left arrow and the answer 3.708 appears under the t-value (actually the two-tailed critical t-value) column.

Difference between AMS and NIST Expanded Uncertainty

There is no difference. AMS is simply one specific special case of the standard National Institute of Standards and Technology ([NIST](#)) methods and consistent [Standard Methods Book](#) methods to calculate “[expanded uncertainty](#).” The reason we call this estimate AMS, instead of expanded uncertainty, is that it is a special case of expanded uncertainty, the case where sample size is always seven (DF = 6) and confidence is always 99%.

Again, the reason for specifying these two conditions is to make this special case of expanded uncertainty as analogous to a MDL (except for the two sided vs. one sided difference) as possible

Just as confidence intervals express the uncertainty about a mean of many different data points, AMS can be used to estimate the interval of uncertainty around each single data point, recognizing that [no single data point is perfect](#).

In cases where MDLs are not optimal, the NPS default suggestion is to calculate an AMS based on [NIST](#) expanded uncertainty using a sample size of 7 and 99% probability. This satisfies two needs at the same time: 1) the need to control measurement sensitivity when in the normal quantitative measurement range and 2) the practical institutional need to have a plus or minus value to put in the STORET “analytical procedure description” text box.

Ideally AMS should be estimated using an extra (not used in calibration) certified reference material (CRM). Such information would be useful for a third purpose, to estimate measurement “accuracy/” Accuracy needs to factor in both bias and precision, and AMS based on seven measures of a CRM could be used as one initial (a bit rough due to the sample size being only 7) estimate for “accuracy.”

Difference between AMS and MDL

To estimate a [MDL](#), one takes the Standard Deviation of 7 measurements of a **blank or other very-low-signal strength sample** times the [one-sided](#) critical t-value for 99% confidence (3.18).

To estimate an AMS, one takes the Standard Deviation of 7 measurements of a **normal sample** (with a signal in the quantitative range) times the [two-sided](#) critical t-value for 99% confidence (3.708).

The following are examples of scenarios in which we might choose to control sensitivity as AMS rather than as a low level detection limit such as a MDL:

In outdoor environments, temperature usually has a strong enough signal to enable us to measure temperature quantitatively in all cases. One would basically never be reporting temperature as below some predetermined [MDL](#) or ML detection limits. In this case, low level detection limits are somewhat irrelevant or at least are not a very relevant sensitivity [QC](#) data quality indicator.

When measuring conductivity in the field or when making **biological or physical** observations in the field, one usually always has a reported value rather than a nondetect, but controlling measurement [sensitivity](#) is still a QC basic that should not be ignored. In these cases sensitivity could be controlled as AMS.

On the other hand in some biological settings (like listening for often faint bird or amphibian calls), low level sensitivity is appropriate and [MDLs](#) should be calculated.

Difference between AMS and AMS+

The difference between AMS and AMS+ is that AMS is based on 7 replicate measures of **one sample**, while AMS+ is based measures of 7 measures of **different samples**, though the AMS+ samples are typically **not separated by much time or space**. We don't call these "field duplicates" (as some do) because we want to make it clear they are not identical samples, though they are often very similar.

Difference between AMS and Precision

Briefly, sensitivity as AMS or AMS+ are similar concepts to [precision](#) and [precision+](#), but AMS and AMS+ are less frequently measured and based on a higher sample size (7 measurements, same as for [MDL](#) sensitivity, to get 99% confidence levels). Precision QC samples are not based on a 99% confidence level but are **more frequently estimated** (usually every 20 measures) QC precision is estimated with a lower sample size (usually two but sometimes three). Precision estimates are usually expressed as [RPDs](#) or [RSDs](#), whereas AMS estimates (like MDLs or MLs) are in original units of measure. For more detail, see [discussion of Precision Versus Sensitivity](#).

Do We Need both AMS and MDLs?

Only in certain cases: If one is measuring some values at very low levels, including some below [MLs](#), and some other levels in the middle of quantitative range above MLs, then one needs both AMS and [MDLs](#) to cover QC sensitivity in both ranges.

If one is only measuring quantitative levels above MLs, then one needs only AMS or AMS+. If one is measuring only very low levels, where most results are below or near the ML, then one need control sensitivity only with [MDLs](#) and [MLs](#).

Why not just use the AMS case of expanded uncertainty for all cases, including low-level detection limits? [NIST](#) has acknowledged that standard NIST methods to calculate "expanded uncertainty" (of which AMS is just one variety) are not applicable for very low (below quantitative-ML detection limits) ranges of measurement (N. Taylor and C. E. Kuyatt. 1994. [Guidelines for Evaluating and Expressing the Uncertainty of](#)

[NIST Measurement Results](#) NIST Publication TN 1297). [MDLs](#) answer a different sensitivity issue (different than zero rather than different from another data point).

Why not just calculate MDLs for everything? The answer is that MDLs are simply not applicable for some scenarios. If no very low magnitude signals are being measured, then sensitivity in that range is irrelevant and zero-value or very low level calibration solutions would also not be relevant to the measurement ranges of interest. Also, in some cases (pH, temperature, and many biological or physical habitat observations, for example), no [blank](#) or other zero-value or extremely-low value calibration solutions (or habitat or biological considerations) are readily available, so one would have a difficult time calculating a MDL even if one tried.

AMS and AMS+ Reporting in STORET

AMS results are not recorded in STORET detection limit fields. Instead, they should be recorded in the STORET metadata “analytical procedure description” text box.

If measurement [precision](#) as repeatability is very good, the AMS result for a single data point might be something like 45.676 plus or minus 0.003. In the more common (for field monitoring) scenario where sensitivity is not that good, the result for a single point might be reported as something more like 50 plus or minus 30.

Either way, the result can be entered into STORET in the plus or minus field for “precision” in the CHEMICAL DATA RESULT ENTRY BOX. Bounding uncertainty in this way is a more modern and defensible alternative to using rounding rules to decide how many significant figures one should carry in final result.

Can one list both low level detection limits and alternative measurement sensitivity for field measurements? Yes, as explained above, there are separate places in STORET to put both results. Whenever one may encounter very low concentrations in some cases and higher levels in others, it would be optimal to do both.

Is it OK to use the lower end of the applicable measurement range (in manufacturer’s specifications) as an estimate of the [MDL](#) and for reporting AMS to STORET? No. This is not ideal and should never be done if the lower end of the specification range is zero. Doing this would never be recommended where true MDLs are needed (for toxics or very low-level nutrients in very pristine lakes). In these cases, proper [MDLs](#) and MLs should be calculated. Zero is never a correct answer.

What if one is always measuring values well above the lower end of the range and does not encounter really bad precision? This is one scenario where an AMS is appropriate. In this scenario, one would expect to be able to meet precision [MQOs](#). Bad precision would be indicated when the field instrument will not settle down on one reading but just keeps changing, even after a reasonable period has been allowed for the instrument to settle down. Really poor precision may be an indication of an instrument or calibration problem, but it can also be a clue that one may unknowingly be measuring values no greater than 2 x an estimated [MDL](#). If it turns out that one is sometimes measuring that close to a [MDL](#), estimating a proper MDL would be appropriate.

Although we recommend they be documented in NPS QA/QC SOPs, MDLs and MLs are only required in NPSTORET if one reports the detection conditions of either “Not Detected” or “Present, below Quantification Limit.” So, if one never encounters this scenario (never being below ordinary detection limits), one need not enter a MDL or ML

into NPSTORET (or in STORET). In this case, one should simply calculate and report a more appropriate type of sensitivity ([AMS](#) or [AMS+](#)) in the QA/QC SOP and in subsequent QC reports and updates).

Censoring AMS Values

Unlike the [MDL](#) or ML, historically data has typically not been censored based on AMS or other expanded uncertainty values. However, as a statistical analysis strategy for looking at a single data point only, one could take the worst case end of the range. For example, suppose the highest pH value considered safe was 9.0 and the only piece of information available was a single value measured was 8.9 plus or minus 0.3. Single values are anecdotal, but one might say the single value could be as high as 9.2 and therefore might exceed the criteria. This would simply be reason to take more measures. If one has multiple values, poor measurement sensitivity simply increases variability and the magnitude of confidence intervals about means or other summary statistics, so there is no need to censor the values.

AMS in Biology and Habitat Observations

Many biological inventory and monitoring projects have not historically estimated measurement [sensitivity](#). However, there is usually no reason why one could not calculate AMS after measuring one sample 7 times (or perhaps have one sample measured by 7 different bio-technicians). It may take some ingenuity in difficult cases. In the case of destructive sampling, it may require a sampling nearby areas rather than re-measuring one identical sample. In the same way we study [precision+](#), we may need to estimate “AMS+” in some cases, where measuring one sample repeatedly is impossible. As long as an AMS+ estimate reflects little variability (below measurement quality objectives, the sensitivity of the measurement process is also well controlled. If AMS+ is high, more study would be needed to find out if the extra variability is from the measurement process or from potential true variability of near but not-identical samples (the + part).

With careful thought, it should usually be possible to develop a common-sense way to adapt the AMS, AMS+ or [MDL](#) functional analogs for various types of biological monitoring. The key is to try do so in a way that “makes sense” while still addressing the issue of logically estimating and controlling measurement sensitivity.

How Often Should AMS or AMS+ be Calculated?

At minimum, AMS (**based on one sample**) or AMS+ (**based on nearby replicates**) should be calculated no less than once a year or whenever methods or instruments change. This is analogous with our related recommendation that [MDLs](#) (a special case of very-low-signal strength sensitivity) be calculated no less than once a year or when methods or instruments change.

AMS+ is an optional QC measure which combines two steps into one, but if the results exceed pre-determined AMS measurement quality objectives, one may then have calculate AMS as well in order to decide if the excess variation was due to true variability of the nearby samples or lack of good sensitivity of the measuring instrument.

Alternatively, networks may specify that **AMS** will be calculated and reported no less than at least once every sampling season or at least as frequently as MDLs. It is important to understand that until a reasonably consistent range, standard deviation, and average for the 7 sample samples is developed, it should probably be done more than once a year.

Like MDLs, AMS should also be re-calculated when something significant in the measurement process changes. For example, if the person measuring, the measuring instrument, or the methods or SOPs change, recalculate AMS to see if measurement sensitivity has changed.

For contrast, one typically estimates QC "[precision](#)" more often (often every 20 samples or once a day) rather than once or twice a year.

Getting seven measures fairly quickly is especially easy when continuous monitoring multi-parameter measuring systems (herein, "**sondes**" for short). See the next section for related tools and discussion.

AMS Tools:

A Northern Colorado Plateau Network (NCPN) staff member has developed some tools helpful for calculating AMS more frequently, including some MS Excel Templates to make calculations of AMS and some other QC metrics easier. The following was suggested by NCPN staff (Dave Thoma, NPS, Personal Communication, 2007):

If users take 7 measurements in a well mixed stream they get the data needed to calculate AMS+. If the stream is not well mixed the standard deviation of measurements should indicate this. So there are two good reasons to take at least 7 measurements at a site. Also, calculating AMS or AMS+ sensitivity on data collected several times during each field season will give a more complete picture of instrument sensitivity over time. Doing this is not as much trouble as some might first think. Sensitivity (and [bias](#)) metrics can be calculated for every field run if field staff simply **take 7 measurements for each core parameter during pre-mobilization final instrument checks (for AMS) and 7 measurements of close but not identical samples at each field site (for AMS+)**. One could also calculate AMS by simply measuring one homogenous sample (or a split sample) seven times. If one had planned to take only one measurement, 6 more could be taken in only about a minute and a half. Why would anyone take just one sample, after all the effort expended to reach field sites? Our experience in well mixed streams is that the following AMS rejection criteria are almost always met in the field in an "AMS+" setting. For individual streams or rivers, less stringent AMS+ criteria could be developed for systems that are not as well mixed.

NCPN AMS Measurement Quality Objectives, Pre-mobilization AMS Stage:

CORE PARAMETER	Sensor AMS MQO (acceptance/rejection criteria)	Optimum Goal/Target (Range) (USGS criteria)
Temperature	± 0.3 ° C	± 0.2 ° C
Specific Conductance **	± 5 µS/cm or ± 3% *	± 5 µS/cm or ±3% *
pH ***	± 0.3 S.U.	± 0.2 S.U.

Dissolved Oxygen ** ± 0.4 mg/L, $\pm 5\%$ saturation ± 0.3 mg/L, $\pm 3\%$ saturation

*Whichever is greater.

** Parameter values (SC and DO) may be affected by 2 to 3% with each degree change in temperature.

*** If after a 2–point calibration and performance of any necessary pH sensor maintenance, error criteria are still not met, a 3-point calibration should be performed before sensor is rejected and replaced.

For more detail, see the latest [Northern Colorado Plateau Network SOPs](#) (intranet site available to NPS computers only).

Using the same thought process that "If one had planned to take only one measurement, 6 more could be taken in only about a minute and a half," then perhaps we should expand our procedures to include the collection of seven AMS measurements at a single point (in the centroid of flow), at each site **on a quarterly basis**, in addition to the cross-sectional measurements to determine [representativeness](#) related to AMS+. The other helpful procedure that we have is our San Francisco Network (SFAN) use of data from our continuous-logging "sondes". Because this equipment is deployed in each watershed for a minimum of two weeks every season, this data can provide any number of data points to determine AMS. Also, we have instituted the additional procedure of taking seven measurements from the deployment cross-section of the stream site both at the time of deployment as well as at the time of removal (in addition to once mid-way through if the deployment lasts longer than three weeks.) This allows for additional checks of AMS+, as well as a correlation for drift and offset for the "sonde" (Rob Carson, NPS, Personal Communication, 2007).

With continuous monitoring sondes, which are more expensive to start with but tend to save money over time, one can easily record a **snap (instantaneous)** measurement at one spot, then move around the stream and record additional snap measurements, or measure at defined intervals in time (or space if a cross section) to get six additional measurements at slightly different locations. Such AMS+ samples would be different but often not much different since they are not separated much by time or space. Such a procedure would give one the seven values one needs to calculate AMS+. The seven values can easily be **automatically recorded** with the sonde system, stored by site and then downloaded back at the office, where they can then be used to quickly calculate AMS+ (or AMS if only one split or homogeneous sample is measured) using the tools discussed just above..

Resolution

In water quality applications, resolution seems most often to refer to some poorly defined notion loosely related to the fineness of the measurement scale. Beyond that, little seems standardized. It is typically not acceptable to use the "resolution specifications" of the manufacturer of a field meter for [AMS](#), a low level detection limit like a [MDL](#), or for [precision](#). There are hints that what many meter, probe, or sonde

manufacturers call resolution in some cases is perhaps some form of uncertainty analysis. However, the lack of consistency between manufacturers regarding how resolution is estimated. This prevents us from calling resolution a synonym for either AMS or some other form of [expanded uncertainty](#).

This lack of consistency might be one reason that the word “resolution” is not typically seen in environmental quality assurance project plans (QAPPs). In any case, “resolution specifications” tend to be optimistic values that do not correspond well to real-world field precision or [sensitivity](#). Sometimes, resolution specifications seem to have been developed at least partly for competitive advantage in ideal lab situations. So, to document measurement sensitivity in the actual environment being monitored, one still has to measure actual field performance for [AMS](#), some other form of [expanded uncertainty](#), or low-level detection limits like [MDLs](#).

The United Kingdom has done more to standardize how resolution of measuring meters is calculated, and they tend to emphasize methods generally consistent with AMS and uncertainty as defined by [ISO](#) and [NIST](#) (UKAS, 2007, [The Expression of Uncertainty and Confidence in Measurement](#)). AMS is simply a [two-sided](#) special case of [expanded uncertainty](#) that is otherwise analogous to (one-sided) MDL detection limit measurement [sensitivity](#), the most common way to report measurement sensitivity in the US.

Until sensitivity methods are more standardized in the US, we recommend that the word resolution not be used in planning water monitoring. Those who have used the word resolution loosely in the US have often been talking about other more commonly understood QC concepts, such as [precision](#), [expanded uncertainty](#), sensitivity, various detection limits, or AMS. Since these concepts are defined in detail separately herein and/or tend to be more universally understood and defined by groups like [NIST](#) and [ISO](#), there is typically no need to address the concept of resolution separately in water quality or contaminants [QAPPs](#) or QA/QC SOPs.

Certain GIS/Remote sensing and non-linear biological categorization applications may be exceptions. In remote sensing, the word resolution is often used for a concept more broadly recognized in other disciplines as sensitivity. In any case, whenever the word resolution is used, how it is estimated should be defined in detail.

IX. Measurement Precision

Precision is simply the variability of the different observations or measurements (of the same thing) when **compared to each other**. Generally speaking, precision checks should only be run on concentrations above the quantitative low-level detection limit, the [ML](#) (3.18 times the [MDL](#), the semi-quantitative detection limit). Any concentrations within 2 times the MDL will have terrible precision (up to 200% RPD, the maximum a RPD can be mathematically), so one does not expect good precision very close to a MDL.

As with many other QA/QC topics, the word “precision” has often been used incorrectly in water quality, contaminants, or statistical literature. It has too often been used for concepts other than variability in measures of one object. For example statisticians have tended to use the word for the size of a confidence interval or for accuracy. Such usages of the word precision should be discouraged to prevent confusion. In concert with [NIST/ISO](#) worldwide scientific consensus definitions, precision is about

variability (scatter), and is clearly not synonymous with confidence, uncertainty in general, or accuracy.

Measurement precision (actually imprecision, but according to tradition and common practice most ignore that) is the variability of repeated independent measures of the same object. [ISO](#) defines precision as “the closeness of agreement between independent test results obtained under prescribed stipulated conditions.” Prescribed conditions are usually either repeatability or reproducibility, as explained in the next section. In any measurement process, precision is not perfect because of random (up and down) imperfections in the measurement process. Unlike systematic error/[bias](#), precision does not depend on the true, right, or expected values, but is just about variability.

Repeatability or Reproducibility Precision?

[NIST](#) has helpfully clarified that the “prescribed stipulated conditions” should include documentation about whether the precision is “precision under repeatability conditions” (where nothing changes) or “precision under reproducibility conditions” (where something changes, see [Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results](#) NIST Publication TN 1297).

Express Precision Results in These Ways:

Unless otherwise justified, for the common QC sample size two (duplicate measures of the same thing) applications, we recommend that precision continue to be reported as a relative percent difference ([RPD](#)) as suggested by the interagency [Environmental Data Standards Council \(EDSC\) 2007 Internet Summary on QA/QC](#). A RPD is simply the absolute value (no negative signs or positive signs) magnitude of the difference between two values, divided by the average of the two values, then multiplied by 100 so that the result can be expressed as a percent (EDSC. 2007. [Quality Assurance and Quality Control Data Standard](#)).

For sample sizes of 3 to 6, precision performance, by common convention, should instead be reported as a relative standard deviation ([RSD](#)). Precision performance could also be expressed as a sample (n-1 version) standard deviation, but this has not been done commonly in water quality or contaminants work so would be strictly optional.

For sample size of 7 (not commonly done for QC control of precision), we suggest that [AMS](#) be reported and that it be called AMS rather than precision.

For sample size of more than 7 (again, not commonly done for QC control of precision), we suggest that [expanded uncertainty](#) be reported and that it be called expanded uncertainty rather than precision.

Regardless of which summary statistic is used to report precision (usually [RPDs](#), but sometimes [RSDs](#)), the QA/QC SOP should state that the raw numbers will also be reported into NPS data bases so that in the future others can look at the precision from different angles. For both STORET and NPSTORET, enter the results for all analyses and their replicates (including duplicates if sample size is 2). The first set of results should be assigned an ‘activity category’ (or ‘type’ for NPSTORET) of 'Field Msr/Obs' for field data or 'Routine Sample' for lab data. The activity replicate(s) of the first activity will then be identified as 'Quality Control Field Replicate Msr/Obs' for field replicates or

'Quality Control Sample-Field Replicate' or 'Quality Control Sample-Lab Duplicate' for lab replicates (duplicates). Consult the STORET and NPSTORET documentation for other types of replicate activities/samples. Calculated precision summary statistics (RPDs or RSDs) can be entered along with each result. How this will be handled could be put in a [separate STORET table](#) (see example herein) or as part of a larger QC table (see next section).

Put Precision Details in a QC Table

The following items to be included in a separate [QC Table](#) in the QA/QC SOP for each protocol:

- A measurement quality objective ([MQO](#)) for precision, usually as a maximum [RPD](#) (RSD is sample size is three).
- Make it clear in the table if precision is being controlled in the context of [repeatability](#) (nothing changes), [reproducibility](#) (something in the measurement process, often a person or instrument, changed), or reproducibility+ (the something that changed included the sample itself). If the details are too complex to put in the table, put the explanation in the precision text discussion section in the QA/QC SOP.
- The data comparability source of the MQO (state, USGS, EPA-EMAP, RCRA, CERCLA, CWA, etc.)? If there is too much information to put this in the QC table, put the details in the comparability section of the QA/QC SOP
- Frequency: Explain how often precision be calculated and reported (for example once every 20 samples). Regardless of the type of precision controlled, precision QC samples are usually performed as duplicate samples every 10-20 samples (or every sampling batch, or every field sampling day, or each laboratory batch).

Standard precision measurement quality objectives ([MQOs](#)) can often be summarized in a [QC table](#), usually along with systematic error, method detection limits, and [blank](#) control MQOs (see Table 9 of San Francisco Network (SFAN) [freshwater quality protocol QA/QC SOP #4](#), for a good example). See also [generic QC MQO tables](#).

For chemical lab measurements, repeatability MQOs are typically used. However, if multiple labs, multiple instruments, or multiple staff members are doing the measuring, it becomes (precision in the context of) reproducibility rather than repeatability.

Precision in the context of [reproducibility](#) is often very relevant for long term monitoring, since there will typically be changes in staff and instruments. Sometimes different staff and different instruments are even used by the same network during one season. In all cases where something in the measurement process changes, precision is controlled in the context of reproducibility. Changes should also trigger efforts to see if the changes introduced new measurement bias that needs to be documented in a [cumulative bias SOP](#).

Specify Precision MQOs as data Acceptance Criteria

Unless otherwise justified, the QA/QC SOP precision section should specify that precision [MQOs](#) specified in a [QC summary table](#) will be used as data acceptance/rejection criteria. Suppose the precision MQO for a particular parameter is that a RPD (of two values, duplicates) cannot exceed 30%. If the RPD exceeds that value, the QA/QC SOP should specify that all values associated with that batch (or that QC sample) will be discarded. Recalibration or other adjustments should then be done until the MQO can be met. When possible, new measurements should then be done to replace the data that was rejected for not meeting precision QC standards (the precision MQO).

If data is to be used for regulatory purposes (and it is always nice to have data useful for multiple purposes, since it is expensive to collect), some states require precision MQOs to be acceptance criteria (See [CA SWAMP](#) example). What will be done if the MQOs are not met? How much data will be rejected? Usually it is all data back to the last time the MQO was met. Will it be all data associated with that batch, that trip, or that day? The rule of thumb is that all data associated with the failed QC performance standard (the MQO) will be rejected.

Historically, many types of biological and physical habitat monitoring measurements have not been associated with attempts to control measurement precision. However, there is growing awareness of the need to do so. One can usually find a common-sense way to control and document measurement precision. Often one can simply measure something twice to get a duplicate answer in either a [reproducibility](#) or [repeatability](#) context. The acceptance criteria MQOs could be loose (even plus or minus 50% or 100% RPD) until more experience is gained, but having no precision MQOs at all is no longer an option ([do something](#)).

Precision+

“Precision+” is our NPS terminology for field duplicates when two or three samples that **are not** exactly the same are taken in close proximity in time and/or space. Since the samples may not be identical, the “+” part of the phrase is a tip off that an additional potential source of variability (true heterogeneity) may be present. This is an extreme case of precision in the context of [reproducibility](#) in that something changed, and is extreme in that one of the things that changed is the sample itself.

In this case, two potential sources of variability are being lumped, lack of perfect measurement [precision](#), and also potential true sample heterogeneity. Another way to say this is that [precision+](#) (simply called field duplicates by some agencies) includes contributors to variability from lack of perfect analytical precision as well as potentially true variability in two potentially different samples.

Using precision+ as the only QC control for precision is an acceptable approach, which combines two steps (measuring instrument [precision plus](#) one rough check on representativeness) into one, but planners should keep in mind that it may eventually trigger additional work compared to the more conventional one-step estimates of measurement precision as [repeatability](#).

We are not suggesting that two different kinds of precision samples should be taken. However, if precision+ is the only kind done, and precision measurement quality objective goals start to be missed, it should trigger another step. This additional step would be to perform some repeatability precision checks (where only one sample is

measured and nothing in the measurement process changes) at the same time one performs precision+ checks, to see which source of variability is most important. If it turns out to be a problem with instrument precision in the context of repeatability, additional calibration may have to be done. If it turns out that the different nearby samples are the main source of extra variability, monitoring teams may have to rethink the overall study design and [representativeness](#) aspects (and possibly then include more randomized samples in time and space to cover a more complete range of conditions in order to better assure representativeness of the [target population](#)

To avoid having to do two types of precision, just doing precision in the context of repeatability is one acceptable approach, though it may result in not finding out more about true heterogeneity of different nearby samples. If one does precision as repeatability only, one is depending even more heavily on survey design controls such as random sampling to obtain representativeness.

Another way to look at this is that [repeatability](#) precision of the measuring instrument (lab or field) combined with the person using it (the two together might be considered the repeatability precision of the measuring system) is often determined by measuring a duplicate on a sample split from one original (single) sample. It might also be a single water sample, well mixed, where one simply inserts the probe twice in fairly quick succession to see the reading is the very close to being the same, or at least within [MQO](#) specifications. The key is that in the case of QC [precision](#), the goal is trying to measure **the exact same thing** repeatedly.

[Precision+](#), for contrast, involves measuring samples that are not necessarily exactly the same, introducing the possibility of additional sources of (true) variability. For emphasis, these additional sources of variability can be related to the separation of the samples in time and/or space, even if the separation does not seem great in the minds of the sampler. Again, if there is some unacceptable difference between two precision+ QC samples, one typically needs to take further steps to determine the source of the difference. For example, temperature can change fairly significantly in stream cross sections and/or according to depth (easily a couple of degrees sometimes), and the change can thus be related to the fact that the variable is really different rather than a reflection of lack of good precision of the measuring instrument. Checking for such differences is the reason USGS suggests preliminary checks first be done to see if centroid stream measurements differ by more than 5% from the composite cross section results ([USGS Field Manual, Wilde and Radtke chapter 6](#)). Such a check amounts to a check on the representativeness of a single measuring point compared to cross section results, but gives no information about representativeness compared to a larger potential target population in space (say farther up or down the river).

Sample Sizes Needed for Precision Estimates:

For typical short term control of precision or precision+, sample size is usually two (duplicates, report as [RPD](#)) or occasionally three (triplicates, for sample sizes of 3 to 6, report precision results as [RSD](#)). Usually the only time that one might want higher sample sizes over time might be to answer the following type of question “Considering QC results from a the last X number of years, what percentage of QC samples exceeded QC measurement quality objectives?” With the exception of USGS WRD, that question

has usually considered less important than the short term data acceptance or rejection question: Should we reject the data, recalibrate, and try again because Precision as a RPD or RSD did not meet the precision [MQO](#)?

Sample size seven QC precision samples from a short time period could be used to calculate [AMS](#), but if sample size were seven, the result would be reported as measurement [sensitivity](#) or [expanded uncertainty](#) rather than as precision. The reason that sample size seven is used for sensitivity estimates is that seven is what has historically been used to estimate low level sensitivity as [MDL](#) detection limits.

Can QC precision replicates be used to increase sample size (other than for [QC](#) issues) for statistical power analyses relevant to the target population? The answer is no, they are used for control and documentation of measurement precision only.

None of the common questions that QC duplicates, QC triplicates, or larger numbers or [QC](#) replicates usually help answer relate specifically to the larger [target population](#) in the environment being measured. So QC replicates are not used to increase sample size beyond one (for statistical power purposes) in trying to answer [common questions](#) about environmental target populations.

One simple explanation for why one should not use QC duplicates, triplicates, or other QC replicates for status or trend power or other statistical analyses is that they are not independent samples in time and space and would thus ‘artificially’ inflate sample size without commensurate addition of information content on true variability (Dave Thoma, NPS, Personal Communication, 2007).

Split Sample Options to Estimate Precision

Some agencies have their own individual terminology for precision samples. USGS distinguishes between “pre-processing split samples” versus “post-processing split samples.” In their Field Manual, USGS also makes the key point that (whether sample size is 2, 3, or more, “replicates are considered identical in composition” ([USGS Field Manual Chapter 4](#)). Well mixed splits would generally be considered one sample, so either way, the result should be reported as precision in the context of [repeatability](#) rather than reproducibility, and sample size relative to target population is one (see discussion in section just above).

Precision Compared To Sensitivity and Detection Limits

As mentioned in the [AMS](#) description, sensitivity as AMS and AMS+ are similar concepts to [precision](#) and [precision+](#), but AMS and AMS+ are based on 99% confidence levels, are much less frequently estimated, and are based on a higher sample size (7 measurements). More **frequently measured** precision QC samples are estimated less often, are not based on 99% confidence, and are based on a lower sample size (usually two but sometimes three). The reason that precision QC samples are done so much more often is to document that the measurement process is remaining “in control.” For contrast, sensitivity is as [AMS](#) or [MDL](#)s is usually calculated much less often, sometimes as infrequently as once per year.

Results are expressed differently too. Precision estimates are usually expressed as [RPDs](#) for QC samples with a sample size two or as a relative standard deviation ([RSD](#))

for a precision QC sample size of three to six. AMS estimates (like [MDLs](#) or MLs) are given in original units of measure rather than as RPDs or RSDs.

MDL low-level detection limits are special-case estimates of low-level sensitivity, and unlike precision (but like AMS) are based on sample size of seven.

A helpful way to understand the relationship between precision and sensitivity is to think of sensitivity is a higher sample size (7) look at precision in certain stated conditions (sample size 7 and certain types of samples). MDL estimates are based on blank or very low-signal strength samples instead of normal-signal-strength samples used to estimate AMS or precision.

[AMS](#) or [MDLs](#) control **measurement sensitivity** less often, whereas precision or precision+ control **measurement precision** more often to make sure the measuring process remains ‘in control’.

X. Measurement Systematic Error/Bias/Percent Recovery

In U.S. water quality and contaminants work, this Quality Control topic has too often been (wrongly) been called “accuracy” by some. Though it has commonly been used in this way in the past, the word accuracy should not be used for the concept of bias. As is understood by [NIST](#) and [ISO](#) and most of the rest of the scientific and engineering world, uncertainty in overall accuracy can only be estimated after factoring in not only systematic error/bias, but also (lack of perfect) precision.

The kind of QC scale **systematic error/bias** being discussed here is the systematic or persistent distortion that would cause each individual measurement to have systematic error in only one direction (usually high or low). On this QC measurement scale of concern, systematic error and bias are usually considered synonyms.

Bias generated from multiple measurements from different samples, the type of bias that would bias a summary statistic such as a mean, say through a faulty study design that does not sure a representative sample from the full range of conditions, is a different topic, related but on a different scale or organization.

In 2006, the [Environmental Data Standards Council](#) (EDSC, an interagency group including EPA, States and Tribes) in its new [Quality Assurance And Quality Control Data Standard](#) guidance; standardized the basic calculation and terminology definitions for bias, stating that [percent differences](#) (PDs, rather than % recovery) should be used. This recent guidance uses the correct ([NIST/ISO](#) compliant) terminology (bias rather than accuracy), but the discussion is short and does not optimally differentiate between the different kinds of bias usually controlled (reference materials, spike expectations, or [blank](#) control expectations). The equation the [EDSC](#) give for **Percent Different** calculations is $PD = [Y - X] / X * 100$, where X is the known (usually “correct” or “expected”) or spiked amount, and Y is the measured concentration.

For contrast, in past water quality and contaminants work, systematic error/bias has usually been expressed as a as % recovery (with the correct or expected answer being considered 100%) based on an interval (such as 80-120%). The new EDSC [Quality Assurance And Quality Control Data Standard](#) guidance specifies that the result should instead be expressed as a % difference compared to the correct or expected answer (say a $\pm 20\%$). The raw values used to calculate these percentages should also be reported as QC results (there is a place for them in STORET and NPSTORET. This is optimal as it

allows one to later look at the results in other ways (such as long term or multi-lab means or standard deviations).

For parameters measured in the field (pH, oxygen, temperature, conductivity, etc.), and for many lab parameters (an exception would be tissue or sediment analyses), a maximum bias (% difference) [MQO](#) of a $\pm 10\%$ is typical (for example see Table A7-1 of EPA 2001. [National Coastal Assessment QAPP](#)).

For quantitative flow measurements, the QA/QC SOP should state how often bias will be estimated, and what maximum % difference will be used as a MQO (such as + 10%, or alternative %?). USGS guidance suggests that bias checks should be performed at least annually or when personnel change and states that QA/QC plans should “Describe what steps are taken to minimize systematic errors. For example, state if field trips are rotated to different personnel every three years or so, or if annually each field trip is performed by someone other than the one who usually performs the trip” (USGS 1995, [A Workbook for Preparing Surface Water Quality-Assurance Plans for Districts of the U.S. Geological Survey, Water Resources Division](#), Open File Report 94-382). When staff doing the monitoring or meters used change, see the [Include a Cumulative Bias SOP](#) section below.

For each measurement done in the field or lab, are the following adequately covered?

- A systematic error/bias measurement quality objective (MQO), such as actual performance should be not worse than a plus or minus 20% percent difference compared to the correct or expected answer.
- Will the MQO be used as a data acceptance performance standard?
- What is the data comparability source of the MQO (state, USGS, EPA-EMAP, RCRA, CERCLA, CWA, etc.)?
- How will systematic error/bias be calculated and reported?
- How often will systematic error/bias be estimated and reported?

Unless otherwise justified, [MQOs](#) should be used as data rejection criteria. For example, suppose the MQO for bias for a particular parameter is that a % difference cannot be worse than a plus or minus 30% of expected. In that case, if actual recovery performance is worse (say plus or minus 40% of expected), all values associated with that batch (usually all data points collected since the last QC samples were performed to check for bias) should be discarded rather than reported into a database as valid results. In such scenarios, recalibration or other adjustments in the measurement process are usually made until the measurement process improves and begins performing within the MQO specifications.

If one value (say water color by remote sensing) is being measured to estimate another value (say chlorophyll a, algae blooms, organic compounds like tannins, or mining wastes), how will bias and accuracy (including a precision component) and [sensitivity](#) be controlled and estimated? Will average observed to expected ratios be used? If so, how will they be used? Will root mean square error techniques be used? How? How will the sensitivity result compare to standard detection limits that use multiples of the standard deviation? If the ground truthing measures include numerous different methods that produce different answers (for example, the numerous different lab

methods for chlorophyll *a*, many of which tend to produce different answers), it complicates deciding what the “right” answer is to compare with remote. In other words, one needs reliable ancillary data to calibrate and validate remote estimates.

Older biological inventory and monitoring data tend to have a total lack of any kind of information on the magnitude of measurement systematic error (bias) or precision at the QC level. In other words, there was not much real attempt to decide what a right answer would look like to be able to estimate bias, or to count or measure something twice to estimate precision. So whereas most past chemical data at least usually has some QC data on % recoveries and for duplicates, no such data exists for much biomonitoring data of the past. Much older biological monitoring data is also largely or totally missing control of QC precision relevant to each data point or minimum detectable differences relevant to multiple data points. This is slowly changing, and one can usually find a common-sense way to control and document measurement bias and measurement precision at the QC level in biological projects, although those working with benthic macroinvertebrates seemed to have progressed farther in this regard than those working with fish. .

One strategy for controlling [bias](#) at the QC level is to consider a senior expert’s answer right or expected (100%) and a rookie trainee’s answer as wrong. This could be done once every 20 samples or even less frequently, but “something” should be done to control bias. As discussed in more detail in [Part B](#) (the longer version), in cases where one is not sure even an expert is “right”, another approach would be to take the maximum difference (delta) between observations as an (maximum, precautionary) estimate for systematic error (bias). This would essentially produce the same comparison couplet as one would get for a precision QC estimate, but bias is usually expressed as a percent difference rather than a relative percent difference (RPD). If one were to express both as an RPD, it would be a bit like doubling the precision estimate, but somehow one has to account for both precision and bias to get at overall accuracy.

In forestry, different terms are used, but bias is controlled. Sometimes, the mean observation minus the known true value is used to estimate bias. Bias in regression analyses can be especially problematic, and rumors indicate outliers are often just (sic, sometimes wrongly) discarded (Chapter 5, Measures and Estimates [in](#) Sit, V. and B. Taylor (editors) 1998 [Statistical Methods for Adaptive Management Studies](#), B.C. Min. For., Res. Br., Victoria, BC, Land Manage. Handbook No. 42.).

Include Calibration Details

In general terms, correct calibration and recalibrating when necessary are important to maintaining data quality in general. EPA considers some calibration issues QA and others QC (for details, see EPA. 2002, Guidance for Quality Assurance Project Plans (QA/G-5, [EPA/240/R-02/009 December 2002](#)).

Calibration is discussed herein as part of the measurement [bias](#) QC section, since maintaining calibration is so important to preventing biased results.

If a field instrument is calibrated in the lab before going to the field, whenever possible, at least one quick final calibration check against a known-value standard (in the range of the values to be measured) should also be performed in the field before measurements begin. This is ideal because an instrument can go out of calibration in

transit, after bumping around and undergoing temperature changes and other stresses on the way to the site, whether the transport is via a truck, backpack, or boat.

For NPS VS monitoring protocols, instrument calibration details should be included either in the QA/QC SOP or in a separate calibration SOP. The need for calibration details is spelled out in the VS generic checklist (“[Checklist for Review of Vital Signs Monitoring Plans](#).”).

If these details are somewhere besides a QA/QC SOP or a separate calibration SOP (perhaps in the field or lab method SOPs), there should be a “point to” in the QA/QC SOP so that the reader will be able to find them. Often an “additional” calibration step should be made in the field to make sure instruments have not fallen out of calibration during transport. Field calibration guidance and other general calibration guidance can be found in [Part C](#) of this guidance.

Blank Control Bias (usually applicable to chemical lab work only)

A blank is typically a sample that is intended to be free of the analytes of interest. Therefore the concentration therein should be less than the [MDL](#) (detection limit documented to be greater than zero). Blank QC samples are tested in the lab to see if they have been contaminated with the analyte during collection, handling, and processing steps. Contamination can result in a positive bias in the reported concentration.

Again for emphasis, blank results are of concern as possibly causing positive bias if the difference between the value obtained by the measurement of the blank sample and the expected value (typically no greater than the MDL) is more than zero. If the result of the measurement of a blank is less than a [MDL](#), then no contamination of the blank is suspected, since we are only confident that a recorded value is less than zero when it exceeds a MDL.

The above is usually consistent with most other federal and state agencies. For example for CERCLA work EPA specifies that “For all blanks, enter the concentration if the absolute value of the concentration is greater than or equal to the appropriate MDL” (EPA. 2004. [Contract Lab Statement of Work for Inorganic Analyses](#)). However, blank control requirements can vary somewhat between various federal and state agencies, so check to see what comparable data sets use.

Unless otherwise justified, for lab chemical measurements of toxic chemicals, metals, pesticides, or nutrients, [MQOs](#) for blank control shall be listed in the QA/QC SOP. QC requirements and frequency of testing blanks shall be no less stringent than State requirements. Depending on whose data will be used for comparison, NPS monitoring may also want to be sure that blank control be no less stringent than that of another federal program whose data will be used for comparison. In other words, we would usually want to standardize not only MQOs but also the frequency of testing blank QC samples, and the type of blank (field, lab rinsate, etc.) with the other agency having considerable data.

Unless otherwise justified, if only one type of blank control sample is to be used, it should go through all field and lab processing steps. If contamination of the blank is found, more work can then be done to find out where the contamination may have been introduced in various field and lab processing steps. Terminology varies, but the concept

is that if only one blank is used (USGS requires more) it should at least go through all field and lab processing steps

For each chemical measurement done in the lab, are the following adequately covered?

1. A blank control measurement quality objective ([MQO](#)), if applicable.
2. What types of blanks will be controlled (trip blanks, lab blanks, etc.)
3. Will the MQO be used as a data acceptance performance standard?
4. Will data reported be adjusted by adding concentrations found in blanks? If not, how will blank control be accomplished (reduce contamination and re-run the samples?).
5. What is the data comparability source of the MQO (state, USGS, EPA-EMAP, RCRA, CERCLA, CWA, etc.)?
6. How will blank control be calculated and reported?
7. How often will blank control be estimated and reported?

Biological inventory and monitoring projects have not historically done blank control. However, if the scenario of wrongly assigning a number value when the true value is zero seems likely, it might be possible to develop a common-sense way to control bias from blanks.

NON-QC SOPS RELATED TO QA/QC

XI. Include a Data Analysis SOP

What are there recommendations for routine data summaries and statistical analysis to detect change? One needs to decide what statistical analyses will be used to analyze the data before designing monitoring details, to be sure that sample sizes and other details are adequate for the analyses being planned.

How often will reporting and trend analyses be done? Does this SOP or the protocol narrative describe the frequency of testing and review of protocol effectiveness?

Statistics that are normally used were first discussed herein above in the sections on [initial statistics to be used](#) and the [size of differences \(MDDs\)](#) that one needs to be able to detect, statistical [power](#), [sample size calculations for mean differences](#), and sample sizes needed for [proportions](#).

The data analysis SOP should include a discussion of the data analyses. This should include: 1) statistics to be used, 2) who will do the analyses, 3) how often the analyses will be done, and 4) what is planned to ensure that adequate staff time and project funding is set aside for this very important task. Most of the proposed statistics should be worked out with a statistician before protocols & SOPs are completed.

Confidence Intervals about a Single Mean

Confidence intervals are good statistics to report, but be sure to specify what exact kind of confidence interval will be reported. The ubiquitous t-distribution confidence interval **about a single mean** is often an easy first step and is available as part of the

standard [MS Excel data analysis toolpak](#). Once toolpak is installed, choose add-ins from the tools menu if not already installed) summary statistics (choose tools, then data analysis, then descriptive statistics, then check summary statistics and then check confidence interval of the mean. The result given is actually the [half-width](#) of the t-distribution parametric confidence interval (the half width is the part of the interval on either side of the mean). However, there is no similar very easy calculation of the confidence intervals of the difference between two means (topic of next section) in MS Excel, and one has to enter the formula manually.

However, in an example of why one has to be careful to get things exactly right when using software, if one uses the “confidence” function in Excel, one instead gets an answer for the half-width of a two sided **Z-distribution** about the mean, and the Z-distribution is the wrong one to use for a small sample size (use the t-distribution for sample sizes less than 30). So make sure you understand and then specify what kind of confidence interval is being reported by the software used. The most common convention for two-sided confidence intervals about a mean is usually to report the mean plus or minus the confidence interval magnitude **on either side of the mean** (the half width on each side of the total interval). To avoid any possible confusion, make it clear which kind of interval is being discussed, especially whenever the confidence interval being reported is really the full interval width rather than the (more common) half-interval width on either side the summary statistic.

Confidence Intervals about the Difference between Two Means

This is a different topic. Confidence intervals around the difference between two means are discussed in Zar (1999), McBride (2005, [Statistics text book](#)) and Helsel, D.R. and R.M. Hirsch. 2002. [Statistical Methods in Water Resources](#)).

A few basics on this “different” kind of a confidence interval (between two means) are presented for comparison:

After properly calculating a 95% confidence interval around the difference between two means, one can see whether that confidence interval includes 0. Doing so is exactly equivalent to conducting a t test with $\alpha = 0.05$. If the calculated confidence interval includes 0, then one concludes there is no significant difference. If the confidence interval does not include 0 then one concludes there is a significant difference. These conclusions will be exactly the same as those from a t test using the same dataset. That's because both procedures (the t test and the confidence interval for the difference between two means) make use of the same statistic, the standard error of the difference between the two means.

There are directions and discussion on calculating a confidence interval for a difference between two population means given on pages 366-367 of Elzinga (Elzinga et al. 1998. [Measuring and Monitoring Plant Populations](#)).

Upper Confidence Interval Limit on a Single Mean

One can get an upper confidence limit on a mean from a two-sided confidence interval, but if the only question one needs to answer is only one-sided (for example, does the mean exceed a water quality standard), a one-sided confidence interval is more appropriate.

The classical (parametric, t-distribution) computation of the 95% upper confidence limits (UCL95) is a t-interval, $\bar{x} + t(0.95, n-1) * s/\sqrt{n}$, where \bar{x} is the sample mean, s is the sample standard deviation, $t(0.95, n-1)$ is the [one-sided](#) critical t-value when alpha is 0.05, $n-1$ is sample size minus 1 (DF), and n is the number of observations. However, this formula assumes either that the data follow a normal distribution, or that we have a lot of data. The minimum sample size needed to use this formula with non-normal data increases as the skewness of the data increases ([Oct06_UCL.pdf](#)).

If park waters are influenced by federal RCRA or CERCLA sites, monitoring planners might want to become more conversant with EPA's default statistical guidance (including information on UCLs):

Box 5-1 in EPA 2000. [Data Quality Assessment guidance for the statistical evaluation of investigative data](#). Practical Methods for Data Analysis, EPA QA/G-9

A statistical program in use in EPA to judge compliance with water quality standards or other benchmarks such as those in use at RCRA or CERCLA sites is Pro UCL, an Excel based add-in macro available on the internet ([Statistical Software ProUCL 4.0 for Environmental Applications for Data Sets with and without Nondetect Observations](#)). It provides options for calculating upper confidence limits (UCLs) in many different ways. One strategy is to try several and pick the result with the longest confidence interval to be precautionary since not all work well at smaller sample sizes:

1. the central limit theorem (CLT) based UCL,
2. modified-t statistic (adjusted for skewness) based UCL,
3. adjusted-CLT (adjusted for skewness) based UCL,
4. Chebyshev inequality based UCL (using sample mean and sample standard deviation),
5. Jackknife method based UCL,
6. UCL based upon standard bootstrap,
7. UCL based upon percentile bootstrap,
8. UCL based upon bias - corrected accelerated (BCA) bootstrap,
9. UCL based upon bootstrap-t, and
10. UCL based upon Hall's bootstrap.

Many in RCRA and CERCLA use this software to deal with site data that will be used to develop statistical Exposure Point Concentrations (EPCs) for a risk assessment performed using EPA's RAGS [[Risk Assessment Guidance for Superfund \(RAGS\) Part A](#)] Risk Assessment Guidance for Superfund).

UCLs can also be calculated with various statistical software programs including "R" and even MS Excel, but not all such options are user friendly.

How should nondetects be used in the estimation of one-sided upper confidence intervals (UCLs)? As summarized in the Practical Stats Newsletter ([Oct06_UCL.pdf](#)), Singh et al. found that “the nonparametric Kaplan-Meier (KM) methods consistently produced the best estimates of the UCL95. Maximum likelihood methods did not provide good coverage for smaller sample sizes or for highly skewed data (so KM would be better than the lognormal MLE recommendation of Frome and Wambach, based on their findings). Probability plot (robust ROS) methods did not work as well as KM - though still much better than substitution. The authors tested several ways to compute the confidence bound around the K-M estimate of mean, and found four to work well: percentile bootstrap, bias-corrected percentile bootstrap, the t formula (using KM estimates of mean and standard deviation), and the Chebyshev formula. The best performance among these four changed with data characteristics -- read their report to fine-tune when to use each of them” (Singh et al 2006. [On the Computation of a 95% Upper Confidence Limit of the Unknown Population Mean Based Upon Data Sets with Below Detection Limit Observations](#), EPA/600/R-06/022).

Also as summarized in the Practical Stats Newsletter ([Oct06_UCL.pdf](#)), one can compute the UCL95 for nondetect data using the percentile bootstrap Kaplan-Meier (KM) method with the KMBMean macro. Nondetect-friendly [Download](#) software is available for SAS, Minitab, and “R” software. NPSTORET transforms nondetects with KM and then calculates certain summary statistics, but confidence intervals are not currently among the statistics that can be calculated with censored data in NPSTORET.

To conservatively estimate optimal sample sizes needed for calculation of an upper confidence limit, see section on [sample sizes needed to estimate the difference between a mean and a water quality standard](#).

Confidence Intervals around Differences between Medians

These confidence intervals are calculated in a different way. Nonparametric confidence intervals about the difference between two medians are discussed in section 5.4.2 in Helsel and Hirsch (Helsel, D.R. and R.M. Hirsch. 2002. [Statistical Methods in Water Resources](#). US Geological Survey Techniques of Water Resources Investigations).

Bacteria and pH Statistics, a Special Cases

Analyzing pH and bacteria data: Be aware that these are reported on a log scale already (actually the negative log for pH). Also keep in mind that the means of the logs will be a bit different than the means for other parameters, since means of logs are essentially geometric means not normal arithmetic means. Also keep in the mind that variability (as expressed by a standard deviation) will usually appear to be lower simply because one is already on the (less variable) log scale. So don't conclude that variability of pH is much lower than other parameters, when the only reason is that pH is on the log scale and the other measures are not. Also be aware of back transformation bias issues. On the other hand, t tests can normally be used with pH and other log transformed data, as the data is already transposed into a more normal distribution type scale. For more details, see the longer version of [Part B](#) and the following internet references:

1. USGS 2006. [National Field Manual section on pH monitoring in general](#).
2. M.D. Mattson. 1995. [Comments on Calculating pH Statistics](#). EPA. The Volunteer Monitor, Vol. 7, No. 2, Fall 1995.

Short or Long Term Trend Analyses?

Interpretation of significance levels of trends can be complicated, and many recent environmental trend analyses have failed to properly account for very long term persistence of trends. For example, is what we are seeing in a given data set possibly a short term up-trend within a longer term down-trend (Cohn, T. A., and H. F. Lins. 2005. [Nature's style: Naturally trendy](#), Geophys. Res. Lett., 32, L23402)? Nevertheless, it is still valuable to document trends, and it is also useful to think about how trends will be analyzed statistically before monitoring begins. The next few sections discuss important considerations when analyzing water quality or other aquatic data for long-term trends:

Consider Diel Differences in Trend Analyses

This topic was first introduced in the section on [representativeness](#). For trend detection for many parameters subject to large [diel](#) swings, one potential strategy is to stratify by time of day (hours after sunrise or before sunset) to try to take out most of the diel variability. The reason for trying to reduce variability is to enable detecting of trends of a magnitude of concern within a reasonable period of time. Most (shallow, sun driven) water-column parameters show diel variability in certain types of locations, especially pH, oxygen, temperature, chlorophyll, many dissolved metals, and nitrates. One sampling strategy that could be stated in protocol narratives is that sampling will be done on a diel basis at first and then later done in restricted periods of time to either get variability down or to capture worst-case time periods. Networks could also try weighting data by time of day, similar to what is done for flow weighting (see next section).

Stratify or Weight by Flow or Water Level Before Trend Analyses?

Considering flow can often help explain why patterns of changes in magnitude and/or changes in variability are happening. The concentrations and loads of many water column parameters (including Total Phosphorus and other parameters that tend to bind to suspended particles) are driven strongly by flow conditions. Although the concentration of extremely water soluble constituents can be a bit less influenced by flow than is the case for some other constituents, it is hard to find water column constituents never influenced by flow rates, especially high flows or first flush (rising limb after a dry period) flows. Therefore, one often has to factor in flow into trend or data interpretation analyses. This is often done in USGS by weighting data for flow. For example, in a recent NAWQA summary, the flow-weighted mean concentration in milligrams per liter (mg/L) was calculated by dividing the total load over the estimation time period by the

total stream-flow (USGS NAWQA 2006. [Nutrients in Streams and Rivers, 1992-2001, Scientific Investigations Report 2006-5107](#)).

The other common way to try to reduce excess variability in concentrations driven by flow is to stratify sampling to include only certain flow conditions (such as only low flow, late summer index periods) to minimize excess variability caused by flow changes rather than a true trend over longer periods of time.

The water level of lakes, ponds, and wetlands can also influence concentrations of water column constituents (evaporation can concentrate pollutants or ions).

One reason we strongly suggest that at least STORET [qualitative flow](#) conditions (dry, no flow, low, normal, above normal, and flood) be recorded in all aquatic Vital Signs monitoring, even if no attempts are made to obtain quantitative flow, is that such classifications are better than nothing and may help in subsequent data interpretation and statistical analyses (for example chemical results associated with low and/or normal might be compared with results classified as above normal and flood flows).

If the result is dry or no flow, quantitative flow measurements are of course not possible, but recording the conditions may still prove helpful to future data users. On the other hand, qualitatively distinguishing between normal, above normal, and flood conditions can be problematic (associated with poor precision reproducibility or repeatability).

Quantitative flow estimates would therefore be superior, even if only the less precise float methods are used. However, no matter how the quantitative or qualitative flow estimates or categories are obtained, the data should be accompanied by QC precision and bias measurement performance results, so that the quality of the data and precision of the categorical groupings can be assessed.

Even quantitative data is also sometimes eventually split into categories before data interpretation or analyses steps. For example, a USGS nationwide approach to show hydrologic variation, involves ranking all annual stream-flow data and dividing in into four “quartiles.” The highest 25 percent of the flow years are classified as “high flow” years, while the lowest 25 percent are considered “low flow” years. The middle range is classified as “normal” and more accurately represents the range of normal conditions than a single figure (K. Blankenship. 2001. [USGS refines long-term flow estimates for Chesapeake](#)).

Consider Phenology Factors in Trend Analyses

Both biological and physical [phenology](#) aquatic indicators, many of which are related to flow, water level, or potential climate change issues; are gaining more interest for future status and trends monitoring. Such indicators could include indicators like the number of ice free days, first date of either ice-free or ice breakup, number of days of flow per year, date or season of first onset and first stoppage of flow, number of days per year a wetland or seep is wetted, dates of onset or length of periods of fish migration or spawning periods, length of algal bloom or algal growing periods (a bit like growing season in terrestrial habitats), date of mayfly hatch (usually temperature related), date of onset of spring high flows, date of end of spring high flows, etc. Many of these indicators may have value related to climate change factors that can drive ecological or biological results.

Once logical categories are defined, there are a large number of parametric and nonparametric statistical procedures available to analyze any resultant **categorical data** (from either qualitative or quantitative data) or **continuous data** (from quantitative flow, see Helsel, D.R. and R.M. Hirsch. 2002. [Statistical Methods in Water Resources](#). US Geological Survey Techniques of Water Resources Investigations.

Consider Seasonal Differences in Trend Analyses

Unless one is sampling in a narrow index window of time (say once a year in late summer only, within a restricted range of water temperatures), to detect trends in a reasonable period of time, one has to do something to take out bias and variability changes that relate to seasonal differences.

In USGS settings, short-term (including perhaps 50 years) trends are most commonly documented in water quality with seasonal Kendall nonparametric tests for trends after one has assured adequate sample sizes.

For lake monitoring, some monitoring groups have tried to factor in seasonality by using parametric regression techniques distinct from simple linear regression in that the data is de-seasonalised (G. Barnes, 2002, [Water Quality Trends in Selected Shallow Lakes in the Waikato Region](#), 1995. Environment Waikato Technical Report 2002/11). The same method is one of the methods used in Lakewatch programs in various parts of the US, including Florida (Univ. Florida [2003. Lakewatch](#)). To be conservative, Florida Lakewatch volunteer monitoring looks at trends using multiple angles and tests (always a good idea) and only calls trends if multiple methods indicated a trend. [Lakewatch software](#) is available (no endorsement implied; we have not tried it).

The deseasonalised Lakewatch parametric tests are mostly based on the work of Noel Burns, who has clarified that he does not emphasize power but instead simply looks at large sample sizes (200) and controls alpha at 0.05. He also uses multiple lines of evidence (chlorophyll, Secchi Disk, TP, TN) to see if most point in the same direction (towards or away from eutrophication). Burns evaluates the coherence of the trends in the 4 variables and translates this coherence, or lack of it, into the probability of change in a lake. Burns also emphasizes looking at the data from different angles (original data and various statistics or transformations) and particularly the seasonal pattern, which can be quite different for each variable and for different lakes. Once the coherence factor between the 4 variables has established whether a lake is changing, a regression relationship between the annual trophic level values for each of the 4 variables and year enables the rate of change in the lake to be determined by LakeWatch (Noel Burns, EarthSoft Consultant, Personal Communication, 2005).

Although I have not seen direct comparisons done, I would expect the deseasonalised method above to produce reasonably similar results to the seasonal Kendall Test (Graham McBride, NIWA, New Zealand, Personal Communication, 2006).

Too Many Choices for Trend Analyses?

For beginners, a thoughtful plain language primer on trends in outdoor environments is [C.A. Stow, et. al.1998](#). (op. cit.). Among the thoughts therein:

1. The choice of trend detection software can be overwhelming because there are so many choices, but this choice is secondary to other factors, like getting enough samples over a long enough period of time and choosing variable with high signal to noise ratios (which usually means ones with less variability, since we cannot manipulate the signals in outdoor environments).
2. Pseudoreplication is not always a fatal flaw in site-specific studies where the questions relate to the site itself rather than many similar sites or many similar stressors. It all depends of the questions. If the questions are not about the effects power plants in general, but about the effects over time of one specific power plant, then pseudoreplication is not such a key issue.
3. Measurement uncertainty (Stow calls it measurement error) can greatly increase variability and therefore inhibit us from finding true trends of importance. Solution: pick variables where measurement error (lack of perfect precision, presence of bias etc.) do not greatly increase variability.
4. Some past publications have argued that autocorrelation and deviations can be relatively minor issues for a realistic limnological time series. Again, evenly spaced samples with large sample sizes can overcome some complications.

Pseudoreplication Issues

A beginner's plain-language explanation of pseudoreplication issues is on the internet [R.J. Irwin and L. Stevens. 1996. [Pseudoreplication issues versus hypothesis testing and field study designs](#). *Park Science* 16(2): 28-31].

These subjects are actually quite complex, and the above plain-language summaries by Stow and Irwin (just above) are not sufficient for those desiring an in-depth understanding of pseudoreplication. Statisticians and other experts would argue for more details and a bit more careful wording due to more recent understandings and clarifications, many of which have been argued back and forth and refined in recent years. On the other hand, many of these details are not easily summarized. See [Part B](#) (heavy) for more detail and some recent publications on pseudoreplication. One example of more careful wording:

If one is not trying to extend the domain of inference beyond a specific site that was sampled through time, then investigators should clearly state that the results and conclusions from their work apply ONLY to that single point, and may or (much more likely) may not reflect results from a more spatially extensive area. Spatial pseudoreplication typically becomes a more-likely problem when the procedure by which one decides which sites to sample from the domain of interest is not implemented in a probabilistic fashion (e.g., randomly, stratified randomly, systematically). Once one has settled spatial issues, one also has to pay attention to temporal issues to make sure that monitoring designs match the question(s) being addressed and the intended domain of interest (Erik Beever, Great Lakes Network, NPS, Personal Communication 2007).

Keep it Simple with Time Period Tests for Step Trends?

Although some of the trends analyses discussed above are complicated, along the lines of Stow's thoughts above, the choice of the method may be secondary to other factors. For example, keep in the mind that if there is evidence of a step trend, there is nothing wrong with using a properly framed hypothesis test (AS ONE PRELIMINARY LINE OF EVIDENCE) to contrast combined data from one time period to another time period. This could be comparing one year to the next. It could also be data from one combined 30-year period (say before a known or suspected step-change event) to the next 30 year time period (say after the step change). Such a procedure might be logical before and after a major known change (for example, the cessation of a large industrial discharge into a small stream, or the time when a major pesticide stopped being used).

Keep in the mind that if there is evidence of a step trend, there is nothing wrong with using a properly framed hypothesis test AS ONE PRELIMINARY LINE OF EVIDENCE to contrast combined data from one time period to data before the (sometimes noticeably abrupt) step change to another time period (after the step change).

Always Try Simple Exploratory Data Analyses:

When planning monitoring (or subsequently fine-tuning monitoring designs over the years in an adaptive manner), there is no substitute for understanding as much as possible about variation in time and space, and getting whatever hints one can by plotting the available data on the X axis against various time (multi-year, within-year, and [diel](#), within 24 hours) and place scales on the Y axis. That is one reason plotting data should typically be part of exploratory data analysis (EDA). If one sees a strong hint of a step trend (a major jog in the plot line) at a single fixed long term monitoring site, then perhaps lumping time periods before and after the jog (step) and doing both the paired version of sample size estimators and the paired versions of hypothesis tests is appropriate. One might also look closer to see if there was logical event that corresponded to the timing of the jog in the line, such as some change in the measurement process (see [Include a Cumulative Bias SOP](#) section below) or a major reduction in point source pollution.

Again, even those who plan to do complex time series (repetitive measure) analyses including trend tests often do common sense checks somewhat similar in theory to a t-test or paired t-test as a part of basic functional data analyses, especially as part of the data analyses step following a data summarization step

Regressions in Trend Analyses:

Although doing regressions is often a first impulse when looking at possible trends, regressions are not used to analyze water quality trends as often as other techniques (notably Seasonal Kendall Tests). This is partly due to the fact that often water quality trends are not linear, and partly since there are often alternatives for trend analyses that are more commonly used. In fact, although all three of the statistics texts listed below this section mention regressions, only the EPA document (the one aimed

least directly at water quality-specific issues) spends much time on regressions in a trend analysis section.

However, since regressions are used partly to define relationships between variables, once seasonal, [diel](#), and/or flow factors have been considered or weighted, there is nothing wrong with using regressions as one additional angle to look at trend issues.

Those deciding to do regressions in testing scenarios in association with trend data (is the slope different from zero, etc.), perhaps as one additional line of evidence in addition to Seasonal Kendall tests when trends are strongly thought to be linear, should think about minimum sample sizes needed to cover the full range of conditions. In addition to the EPA sample size calculators for trends discussed, above, there are also various generic sample size calculators aimed at more complex regression topics, including hierarchical multiple regression analyses. These topics are too complex to discuss briefly herein, but try to only use sample size calculators that include inputs for beta and alpha rather than alpha alone.

Missing Values, Useful Data, and Effective Data

The Data Analysis SOP should detail how imperfect data can be and still be used in data analysis or to meet [completeness](#) goals. Unless otherwise justified, data that have not met QC measurement quality objectives for [precision](#), [bias](#), and [sensitivity](#) are not considered useful and are not included in quantitative statistical analyses. The same is true for: 1) data below qualitative detection limits ([MDLs](#)), 2) data between MDLs and MLs (see detection limit discussion further below herein for exceptions), 3) data associated with holding times that have been exceeded or where preservation requirements have otherwise not been met, 3) chemical concentrations where improper containers were used, 4) data beyond minimum and maximum plausible values (checked via range sensibility checks). If some of these will be considered OK for qualitative data analysis, provide the rationale in the data analysis SOP.

How will missing values be handled? This topic is highly related to completeness goals, but decisions hinge not only on what % (like 15%) can be missing, but also on whether or not the missing data is from a critical class of data. For example, suppose the question is “What is the annual temperature?” If all 15% of the data missing are in the coldest part of the year, it would tend to bias the answer.

When using the seasonal Kendall test for trends, an allowance for missing data can be made. In fact, non-parametric tests are sometimes chosen because 1) they are not affected when the distribution of data is not normal, 2) are insensitive to outliers, and 3) are less impacted (or not impacted) by missing or censored data (Harcum, J.B., J.C. Loftis, and R.C. Ward. 1992. Selecting trend tests for water quality series with serial correlation and missing values. *Water Resources Bulletin* 28(3):469-478). The decision of what percent of the data can be missing and still pass [completeness](#) goals should include a common sense check relative the questions to be answered and the statistics to be used.

Include this topic in discussions with your applied statistician, since imputation options tend to be complex. See [Part B](#) for additional discussion.

Part B also contains generic definitions for useful data and effective data. Depending on the questions and project data quality objectives, data from screening methods can be “effective” even it was not collected by the most precise or accurate methodology (see detailed discussions of effective data in Part B).

Although the data analysis SOP should cover the basics of what will be done with missing data, more detail, along with how values below detection limits (see discussion further below related to low level detection limits) will be handled, should be covered in the QA/QC SOP.

Useful References for Statistical Analyses:

The first three cover water quality and contaminants scenarios and also cover many trends-related issues, including Kendall and seasonal Kendall tests, Sen's Slope Estimator, other options for assessing trends, and autocorrelation:

- 1) Graham McBride. 2005. [Using Statistical Methods for Water Quality Management: Issues, Options and Solutions](#). Wiley, NY, 313 pp.
- 2) Helsel, D.R. and R.M. Hirsch. 2002. [Statistical Methods in Water Resources](#). US Geological Survey Techniques of Water Resources Investigations).
- 3) EPA 2000. [Data Quality Assessment guidance for the statistical evaluation of investigative data](#). Practical Methods for Data Analysis, EPA QA/G-9. This one is not focused on water quality or aquatic work but is a useful statistical text on the internet.
- 4) D. Helsel. 2005. [Nondetects and Data Analysis: Statistics for Censored Environmental Data](#). Wiley.
- 5) EPA explains many probabilistic analysis issues at an [EMAP monitoring and design and analysis](#) home page. This is a good reference but probably cannot be used as a stand alone (don't just say analyzed according to EMAP suggestions).
- 6) A recent list of internet statistical calculators in general is in EPA. 2007. [Analytical and Sampling Tools](#). Other EPA N-Step tools cover [change-point analysis, correlation, and regression](#) issues.
- 7) Robert C. Ward, Jim C. Loftis, Graham B. McBride. 1990 [Design of Water Quality Monitoring Systems](#), books.google.com.
- 8) Adaptive Management Statistics (Sit, V. and B. Taylor (editors) 1998 [Statistical Methods for Adaptive Management Studies](#), B.C. Min. For., Res. Br., Victoria, BC, Land Manage. Handbook No. 42.). This entire text book is on the internet and includes an introduction to Bayesian methods and to studying uncontrolled—out in nature--events).

- 9) Various references and discussions available to NPS employees only at the [NRPC water quality statistics Sharepoint site](#).

XII. Include a Cumulative Measurement Bias SOP

Is our internal NPS data from both old and newer measuring systems comparable enough that the different sets of data could be combined for purposes of determining trends or making management or regulatory decisions?

“Do not be swayed by the argument that we cannot change now because ‘we have 4 years of data that will be compromised.’ You are in this for the long haul. If you do not correct mistakes now, 25 years later you will (should) be cursed” K. Burnham. 2004. [Wildlife monitoring: success requires more than a good sampling design](#)).

The SOP should make it clear that data adjustments would be made for trend analysis only rather than before reporting data into long term data bases (the networks or STORET). More detail: It gets tough to estimate the effects of cumulative bias of say eight changes over say 100 years of monitoring, and for data reporting, we would not usually want all the data normalized to methods from 100 years ago, since the older methods may be more biased and/or less than precise than the newer methods, So the real utility of keeping track of cumulative bias changes, the focus of this section, is to try to make it easier for future data users to get all the data into comparable units for long term trend analysis. In the past it has been difficult for data users to determine if some of those jogs in the trend line might be explained by method changes instead of true changes in environmental variables.

A good example of a good cumulative bias SOP is the [SOP 6 of the Northern Colorado Plateau Network Freshwater Protocol](#) (intranet site available on NPS computers only).

[NIST](#) suggests adjusting data to correct for bias, but chemists have ordinarily not done this because sample sizes of the bias comparisons have usually been too small to get a good estimate of the magnitude of the bias. Data for reporting into database such as STORET would not be adjusted, and even data used in trends reporting would not be adjusted unless:

1. There is a strong indication of which set of measurements were inferior (the old or the new), and
2. If sample size was only one (the common case for recurrent QC bias checks -- % recovery bias -- every 20 samples) sample size is clearly not large enough for a good estimate of the bias. So in the case of sample size of one (one couplet producing one RPD) for routine QC checks, one would never adjust the data, sample size is simply much too small.
3. If sample size was only seven (the case herein when only the observer changed) sample might still be large enough for a really good estimate of the magnitude of the bias. Therefore, if less than 25 couplets of old and new measures were compared (this would apply mostly to the sample size 7-25 comparisons), and IF sample size was adequate to be able to detect at least a 10% (or greater) difference between the old and new means with 90% power and a significance level of 0.10 (as determined with [paired](#)

[sampling sample size calculators](#)) then one might adjust data by the average amount of bias change for trend analysis only, not for reporting into databases such as STORET. Flunking this criterion would be more likely when sample size is less than 25.

4. If more than 25 couplets (old versus new data) were compared, and the RPD exceeded 10%, then the estimates of each mean (old versus new) would be considered reliable to adjust data for trend analysis only, again not for reporting into databases such as STORET.

Why limit both Type I error (the risk of deciding there has been a change when there has not been one) **and** the Type II error (the risk of concluding there has been no change when there has been a change) to the same %? The answer is that we simply want to know whether there has been a change in measurement bias or not, and in this situation, there is no particular reason to be more concerned about one type of mistake than the other. Why use 0.10 instead of 0.05? For this purpose, we simply consider 90% confidence sufficient. If networks want to use 0.05 for alpha to be consistent with tradition, that would be acceptable, but then they should probably also consider making beta 0.05 too, and keep in mind that using 0.05 instead of 0.1 would drive up required sample sizes.

Although many seem to understand that changing observers will sometimes bias qualitative or semi-qualitative eye-ball estimates (such as % embeddedness of cobbles in a river bottom) up or down. Less broadly recognized is that even using:

- 1) what seems to be very similar hardware or
- 2) identical calibration solutions

does not guarantee that changes in instruments will not bias the newer readings up or down compared to the older readings using the other instrument. In one example, oxidation-reduction-potential (ORP) values measured by instruments of two manufacturers were very different even though a virtually identical probe and the same type of calibration solution were used by both manufacturers. Although the two manufacturers also used similar calibration protocols, they recommended use of two different reference scales (Pete Penoyer, Personal Communication, NPS, 2006). Those who will actually be measuring ORP in groundwater can find more detail in [Part B](#).

[WRD](#) as well as VS/NRPC Database Staff (Margaret Beer) believe that documenting measurement bias after overlapping old and new methods is important enough to warrant its own SOP. It would also be acceptable for the monitoring network to choose to document this type of information in the Data Analysis SOP, if a good rationale for doing so is provided. However, for NPS VS consistency, we recommend that a separate Cumulative Measurement Bias SOP be included.

Either way, the information is important enough for those who will eventually be trying to detect trends that liberal use of “point-to links” should be included in the Data Analysis SOP, so that future data users can find this information.

Method, equipment, and personnel changes are inevitable in long term monitoring. The requirement of overlapping old and new measurement methods is in [Oakley et al.](#)

(2003). However, both this requirement and the underlying reason for it are too often overlooked.

The SOP should detail how long the old and new methods are to be overlapped to determine changes in measurement [precision](#), [sensitivity](#), and (especially) measurement [bias](#). The following text (or something like it and defensible) is suggested for inclusion in a Cumulative Measurement Bias SOP, with crosslink text in the data analysis SOP pointing to the SOP where this may be found:

When the Only Change is a Change in Personnel
Doing the Measurements, Observations, or Ratings:

Single (identical) samples will be measured by old and new personnel at least 7 times when the only thing changing is staff doing the measuring or observations.

When the Change is a Change in Meters,
Measurement Instruments, Methods, or SOPs:

At least 30 overlap measurements will be made when a method, SOP, meter, or measuring instrument changes.

When the Change is a Change in an Indicator
Or in One Surrogate Measure to Estimate Another

At least 50 overlap measurements will be made and results recorded. The bigger the method or SOP change, the more repeat sampling may be appropriate. Some states have gone to great lengths when replacing one indicator with another. For example, Oregon created regression derived equations (based on sample sizes larger than 50) for estimating fecal coliform values from *Escherichia coli* values (or vice versa) after converting to *E. coli* monitoring. They monitored both fecal and *E. coli* side by side for six years before deciding that *E. coli* = about half fecal coliforms for Oregon's rivers and streams and becoming comfortable with dropping fecal coliforms (C.G. Cude. 2005. Accommodating change of bacterial indicators in long term water quality datasets. Jour. Am. Water Resources Assn. 41(1): 47-54). An indicator change is a bigger change than a staff change or a method change.

When the Change is a Change in the Basic Sampling Design

In this scenario, due to changes in the [Target Population](#), one makes changes in the basic sampling design, including changes in the sample frame, or a re-randomization of potential samples from an existing (or new) sample frame. For example, one might be selecting a totally new mix of samples from a new [GRTS](#) draw. One might also decide to re-stratify or make some other substantial change in how samples are selected. In these cases, to correctly determine trends one needs to know if the changes in sampling design changed the results, or if the results changed due to a true change in the target population in the environment. If the sample frame was wearing out so that it was no longer a good

representation of the target population, the change might decrease a bias that was creeping in before the change was made. Unlike the bias that might change due to factors described in the paragraphs above (which relate to each single data point), the type of bias we are discussing here is on a different level of organization (multiple measurements made according to monitoring design).

In this case, at least 30-50 overlap measurements will be made and results recorded. So for example, 30 samples would be measured based on the old sampling design (including the old randomization and stratification scheme), and another 30 samples would be drawn from the new scheme and measured to determine the differences in old way versus new way.

In all of the cases listed above, the following shall be archived in the Cumulative Measurement Bias SOP:

1. The sample size, standard deviation, and average % bias change from the old measurement system to the new, calculated as an average of percent differences (PDs, not RSDs or RPDs). If an initial sample size of 30 was chosen as a starting point (from the three options listed above), one would do 30 comparison measures side by side, calculate a PD for each one, then average the 30 PDs to get an average PD for those 30 comparison pairs (this would be considered “paired sampling”). Each PD change is calculated by subtracting the old measure from the new one, then dividing the difference times the old measure, then taking the result times 100. In other words, $PD = [(new - old) / old] * 100$. By subtracting the old measure from the new one, if the change is positive, the PD calculated value will be positive too.
2. In the case of sample sizes of less than 25 pairs of old and new data, the results of a paired t-test of the differences of the two means, based on alpha of 0.10, power of no less than 90%.
3. The precision as reproducibility or repeatability RPDs or RSDs.
4. Measurement sensitivity, as either a MDL (if some measurements are near or below the MDL) or AMS (if all measurements are well above the MDL or if the MDL is otherwise not appropriate, see separate discussions herein) for both the old and new measurement systems
5. The date that the overlapping measurements started
6. The date that the overlapping measurements stopped (some kind of date is needed since a change in methods may help explain a jog in a trend line, especially if the jog happened just after the change).
7. The date that the average percent difference bias change from old to new measurements was calculated.
8. All paired raw values, should future statistician desire to normalize values a different way when estimating trends.
9. There should be a clear statement of which way the bias went. If the values for the new measurement system are on average higher than those for the old (PDs are on average positive), the bias resulting from the change is positive.
10. Trends are then based on values normalized to the original numbers. If the average PD based on the 30 samples was a plus 5% (the new meter on average read 5% higher than the old meter, for purposes of trend analyses, the new

values can then be normalized to old by multiplying the new values times the calculated fraction of change, so one would multiply the new values times 0.95..

Although the average PD is not a proportion, each data point to estimate the average PD was a proportion and therefore would have ideally been based on a sample size of 25 or above so that each data point that went into the average PD calculation was optimally credible (see separate discussion in Section herein entitled “[Sample Size Needed to Estimate a Single Proportion](#)”).”

If there are logical reasons why the estimate of bias change should be controlled even more tightly than suggested above (for example, the measures are near a magnitude that would bring one is near a collapse [threshold](#) value for a rare resource), there is nothing to prevent a monitoring network from controlling the estimate more tightly. Controlling it more tightly would usually simply require larger sample sizes (sometimes much larger if the variability was high) for the number of overlap measures.

One side benefit of doing overlapping measurements is that one might discover the old meter or method is better and decide not to use the new one. Also, if a change in personnel changes the bias or precision in unacceptable ways, it is best to find that out as early as possible so that additional training or other changes can be made until an acceptable result is obtained.

Ideally, one would also want the magnitude of bias changes (calculated as per above) to be two times lower than minimum detectable differences ([MDDs](#)) calculated for overall [monitoring design sensitivity](#) or [AMS](#).

Is the above too much to ask? We do not believe so, and the cost of not documenting cumulative measurement change bias is an inability to differentiate true environmental trends from measurement or estimation changes. We have seen dramatic examples where one could not differentiate trends from method changes in past data from the last 50 years, and we would like to avoid that with our new monitoring program. Why bother to monitor long term if we don't do it in ways that allow us to document true trends in defensible ways?

Often due to inevitable changes in staff and measuring equipment in long term monitoring, the data are not comparable enough to differentiate trends from changes caused by changes in measurements instruments, staff, or personnel. Therefore, overlapping measures need to be done for a period of time to see if measurement bias has been introduced from the old measurements to the new.

Even volunteer groups are performing these kinds of method change comparisons now. One expert recommended that volunteer groups overlap 30-50 paired observations when changing salinity methods [P. Bergstrom. 2005. Comparing [Four Salinity Methods](#). 2005. *The Volunteer Monitor* 17 (1):21]. The NPS typically does not want to be less rigorous than volunteer groups; 25-50 is often a minimum sample size to estimate many summary statistics (such as proportions and means) well, so unless otherwise justified, overlap at least 30 samples for method or other substantial changes.

Even small changes in measurement bias can accumulate and become significant over time. “Point-to” notes about where such documentation is should also go with the data, as part of metadata notes or introductory notes.

The goal would be for someone 100 years later to be able to discover the effect of the various changes in measurement bias. These are the types of examples that a future

data user would need to discover to enable that user to separate true trends from method changes: 1) 90 years ago there was a method change that resulted +2% change (on identical samples) from the previous method, 2) 80 years ago there was another method change that resulted in another change of +4% from the method used in the previous 20 years, 3) 60 years ago there was another method change that resulted in a new plus 3% bias, and 4) 75 years ago there was another method change that resulted in a -1% bias from the years just before. In this pretend example, unless the future data user could find this type of information, that person might conclude there was a steady upward trend that leveled off a bit at the end of the period, when the only changes were really a series of bias changes caused by method or SOP changes. This issue is important enough for long term monitoring that some redundancy provided by the multiple “point to” links from other places seems prudent.

How the data will be normalized could be handled in either SOP with “point-to” links from the other. FOR PURPOSES OF TREND ANALYSIS ONLY (not for adjusting data before reporting it into a data base such as STORET, usually all data will be normalized to the original measurement method. For example, in 2100, data from 2006, 2020, and 2040, etc. might all be normalized to 2006 data equivalents. If it is determined that the original method used in 2006 was too deficient to form a defensible normalizing starting point for example, measurement [precision](#), measurement [sensitivity](#), and/or measurement [bias](#) were bad or incompletely documented, one option would be to start over with a new normalization point (say 2020 or 2040) after the deficiencies in documentation of measurement performance have been corrected.

In summary, it is suggested that the cumulative results of the bias over the years be detailed in the Cumulative Measurement Bias SOP in each protocol with “point to” hyperlinks from other places people might look, such as the protocol revision log, each field and lab SOP for methods, the data management SOP, the data management section of the protocol narratives and central monitoring plan, the data acquisition parts of the central monitoring plan, the Data Analysis SOP, and the precision and bias discussions in the QC SOP.

XIII. Include STORET Details in a Data Management SOP

An updated starting point for those wishing to become familiar with WRD guidance is the [Water Quality Data Management and Archiving](#) guidance.

QA/QC results should be reported into STORET or NPSTORET as summarized in the Table below (For convenience in comparisons, the following table is in the same order as a similar table explaining the technical detail differences between these metrics in the separate chapter above entitled “INCLUDE A QA/QC SOP):”

STORET/NPSTORET Reporting of QC Measurement Quality Indicators

Purpose	Description	Metric Acronym	STORET Note	NPSTORET Note
---------	-------------	----------------	-------------	---------------

Purpose	Description	Metric Acronym	STORET Note	NPSTORET Note
EPA and State Low Level Sensitivity as Detection Limits (Usually Lab)	Lowest value that can be differentiated from zero	Method Detection Limit (MDL) – for control of very low level sensitivity	<p>Detection Limit:</p> <p>Put MDL in the Detection Limit field (APL2 Result Laboratory Data Entry) for all applicable results. Use the Description field to indicate what type of Detection Limit (MDL) was entered. If a result is reported as <MDL, set Detection Condition to “Not Detected” on R4 Chemical Result Data Entry screen.</p>	<p>Detection Limit:</p> <p>Enter MDL for each applicable characteristic definition on the Metadata Template, Characteristics tab, Detection Limit field. Alternatively (and for MDLs that change frequently), enter the MDL on the Results Template, Detection Limit field. If a result is reported as <MDL, set Detection Condition to “*Non-Detect” on the Results Template, Detection Condition field. In both instances, use the Detection/Quantification Limit Description field to indicate what type of Detection Limit (MDL) was entered.</p>
USGS Low Level Sensitivity as Detection Limits (Usually Lab)	Lowest value that can be differentiated from zero based on long- term data	Long Term Method Detection Limit (LT-MDL) – for control of very low level sensitivity	<p>Detection Limit:</p> <p>Put LT-MDL in the Detection Limit field (APL2 Result Laboratory Data Entry) for all applicable results. Use the Description field to indicate what type of Detection Limit (USGS LT-MDL) was entered. If a result is reported as <LT-MDL, set Detection Condition to “Not Detected” on R4 Chemical Result Data Entry screen.</p>	<p>Detection Limit:</p> <p>Enter LT-MDL for each applicable characteristic definition on the Metadata Template, Characteristics tab, Detection Limit field. Alternatively (and for LT-MDLs that change frequently), enter the LT-MDL on the Results Template, Detection Limit field. If a result is reported as <LT-MDL, set Detection Condition to “*Non-Detect” on the Results Template, Detection Condition field. In both instances, use the Detection/Quantification Limit Description field to indicate what type of Detection Limit (LT-MDL) was entered.</p>
EPA and State Lower Quantitative Sensitivity as Detection Limits (Usually Lab)	Lowest Quantitative Value	Minimum Level (ML) – Values higher than ML are quantitative	<p>Quantification Low:</p> <p>Put ML in the Quantification Low Limit field (APL2 Result Laboratory Data Entry) for all applicable results. Use the Description field to indicate what type of Quantification Limit (ML, LQL, etc.) was entered. If a result is reported as <ML, set Detection Condition to “Present, below Quantification Limit” on R4 Chemical Result Data Entry screen.</p>	<p>Lower Quantification Limit:</p> <p>Enter ML for each applicable characteristic definition on the Metadata Template, Characteristics tab, Lower Quantification Limit field. Alternatively (and for MLs that change frequently), enter the ML on the Results Template, Lower Q.L. field. If a result is reported as <ML, set Detection Condition to “*Present <QL” on the Results Template, Detection Condition field. In both instances, use the Detection/Quantification Limit Description field to indicate what type of Lower Quantification Limit (ML, LQL, etc.) was entered.</p>

Purpose	Description	Metric Acronym	STORET Note	NPSTORET Note
USGS Lower Quantitative Sensitivity as Detection Limits (Usually Lab)	USGS alternative to the ML based on long term QC data and LT-MDLs	Long-term Reporting Level (LRL) – Values higher than LRL are quantitative. (Unique to USGS laboratory)	.Quantification Low: Put LRL in the Quantification Low Limit field (APL2 Result Laboratory Data Entry) for all applicable results. Use the Description field to indicate what type of Quantification Limit (USGS LRL) was entered. If a result is reported as <LRL, set Detection Condition to “Present, below Quantification Limit” on R4 Chemical Result Data Entry screen.	Lower Quantification Limit: Enter LRL for each applicable characteristic definition on the Metadata Template, Characteristics tab, Lower Quantification Limit field. Alternatively (and LRLs that change frequently), enter the LRL on the Results Template, Lower Q.L. field. If a result is reported as <LRL, set Detection Condition to “*Present <QL” on the Results Template, Detection Condition field. In both instances, use the Detection/Quantification Limit Description field to indicate what type of Lower Quantification Limit (USGS LRL) was entered.
Upper Quantification or Quantitation Limit	Quantification refers to the limits of an instrument or analytical process when detecting and/or quantifying a substance associated with a result value. High represents the largest amount of the target substance that could be quantified by the instrument or analytical process; Low (ML or USGS LRL) represents the smallest amount. Values above the minimum and below the maximum quantification limits are reported as valid numeric results.		Quantification High: Put upper quantification limit in the Quantification High Limit field (APL2 Result Laboratory Data Entry) for all applicable results. Use the Description field to indicate what type of upper quantification limit was entered. If a result is reported as > upper quantification limit, set Detection Condition to “Present, above Quantification Limit” on R4 Chemical Result Data Entry screen.	Upper Quantification Limit: Enter upper quantification limit for each applicable characteristic definition on the Metadata Template, Characteristics tab, Upper Quantification Limit field. Alternatively, enter the upper quantification limit on the Results Template, Upper Q.L. field. If a result is reported as > upper quantification limit, set Detection Condition to “*Present >QL” on the Results Template, Detection Condition field. In both instances, use the Detection/Quantification Limit Description field to indicate what type of Upper Quantification Limit was entered.

Purpose	Description	Metric Acronym	STORET Note	NPSTORET Note
Sensitivity (Usually Field, or whenever MDL is N/A)	Determines instrument noise in both directions (up or down)	Alternative Measurement Sensitivity (AMS) – Lowest change possibly real	Entered in the Description field of the Field/Lab Analytical Procedure metadata (P3 Organization Field/Lab Analytical Procedure Data Entry) and/or the Precision +/- field for each result on the R4 Chemical Result Data Entry screen.	Enter alternative measurement sensitivity for each applicable characteristic definition on the Metadata Template, Characteristics tab, Characteristic Description Field. As with STORET, you can also enter the alternative measurement sensitivity in the Analytical Procedure Description field of the Metadata Template, 4. Analytical Procedures tab. Alternatively, enter the alternative measurement sensitivity on the Results Template, Precision (+/-) field for each result as appropriate.
AMS+ (Usually Field, or whenever MDL is N/A)	Includes instrument noise and natural heterogeneity	Alternative Measurement Sensitivity+ (AMS+) – Total variability of close replicates	Enter in STORET as stated above if no other form of AMS is reported.	Enter in NPSTORET as stated above if no other form of AMS is reported.
Precision (Lab and Field)	Variability of repeated measures (precision)	Relative Percent Difference (RPD) – QC Precision Control	Enter the results on two separate activities. First activity category (FA2 Sample Data Entry) is 'Routine Sample'; second activity category is 'Field Replicate/Duplicate'. Compute the Relative Percent Difference, and, optionally, include the Relative Percent Difference in the Comments field on the R4 Chemical Result Data Entry screen for each applicable result.	Enter the results on two separate activities on the Results Template. First activity type is 'Sample-Routine'; second activity type is 'Quality Control Sample-Field Replicate'. Compute the Relative Percent Difference, and, optionally, include the Relative Percent Difference in the Comment field on the Results Template for each applicable result. Alternatively, use the Reports & Stats Template, Statistics tab, Precision Analysis to compute Relative Percent Difference and display these values in a summary report.
Precision+ (Usually for Field Measurements Only)	Variability of repeated measures (precision +) + = potentially some additional true variability (two samples not one)	Relative Percent Difference (RPD) – QC Precision+ Control	Enter in STORET as stated above if no other form of Precision is reported.	Enter in NPSTORET as stated above if no other form of Precision is reported.

Purpose	Description	Metric Acronym	STORET Note	NPSTORET Note
Bias (Lab and Field)	Difference from standard (bias)	Percent Difference (PD) – QC Bias Control	Select the appropriate activity category (e.g. 'Field Spike') on the FA2 Sample Data Entry screen and enter the results. Compute the Percent Difference from the expected value for each result and include the Percent Difference in the Comments field on the R4 Chemical Result Data Entry screen for each applicable result.	Select the appropriate activity type (e.g. 'Quality Control Sample-Field Spike') on the Results Template and enter the results. Compute the Percent Difference from the expected value for each result and include the Percent Difference in the Comment field on the Results Template for each applicable result.
Blank Control Bias (Usually for Lab Measures Only)	Difference between measurement result and blank sample expected result (usually no greater than the MDL)	Percent Difference (PD) – QC Blank Control Bias	Select the appropriate activity category (e.g. 'Field Blank') on the FA2 Sample Data Entry screen and enter the results. Compute the Percent Difference from the expected value (e.g. MDL) for each result and include the Percent Difference in the Comments field on the R4 Chemical Result Data Entry screen for each applicable result. Record both the measured value and the MDL.	Select the appropriate activity type (e.g. 'Quality Control Sample-Field Blank') on the Results Template and enter the results. Compute the Percent Difference from the expected value (e.g. MDL) for each result and include the Percent Difference in the Comment field on the Results Template for each applicable result. Record both the measured value and the MDL.

NPS monitoring networks may want to copy parts (or all) of the table above into their Data Management SOP attached to each protocol.

Documentation and planning in the SOP needs to include matching the network's characteristics/parameters with the official standardized EPA list of 389,007 (as of 12/1/2005) characteristics (found in tblDef_TSRCHAR in NPSTORET ([Storet Characteristics Tables](#))). For questions, contact [Dean Tucker@NPS.GOV](mailto:Dean_Tucker@NPS.GOV)). How the data collected will be archived in STORET and NPSTORET should be detailed in a Data Management SOP.

The Data Management SOP needs to include a data dictionary (DD) that clearly defines what is in each field. Aim for the sweet spot between too long (hard to use) and too short (not clear). This is a hard balance but still worthy of effort. STORET, NPSTORET, the NWQMC 2006 [Water Quality Data Elements](#) (= both data and metadata), and [Part B](#) (the longer version) have all been criticized as being too long. But when their authors tried to remove things to shorten them, the new versions were criticized as being too short or incomplete, a catch 22. Regardless of the difficulty, the network goal is to make the data dictionary clear, but not too long.

End of Part B lite. More detail on each of the topics in Part B lite is found in the long version of Part B at <http://science.nature.nps.gov/im/monitor/protocols/wqPartB.doc>.