3.0 EVALUATING THE ABILITY TO DETECT CHANGE

The ability to detect change is the first step and is crucial to the technical assessment guidance process. Detecting change, or differences in ecosystem responses to CERP implementation, is a function of three highly correlated factors:

1) The expected (baseline) variability of the hydrologic, water quality and/or ecological indicator
2) The strength of the response being measured
3) The power of the sampling design to detect a change in the response that is significantly different from natural or background variability

Detecting change is inextricably linked to measuring variability and the statistical design of the monitoring program. Consequently, a primary goal of the assessment process is to measure the "baseline" variability with specified degrees of confidence so that we can detect changes when they occur. Further discussion of establishing baseline variability and reference conditions is presented in Section 4.

3.1 Purpose

The objectives of this section of the technical assessment guidance process are as follows: 1) describe the experimental design and the supporting scientific rationale; 2) describe the results of the power analysis for the sampling design; 3) determine the minimum detectable difference of the power analysis and its associated confidence and uncertainty for each of the performance measures in the module; and 4) describe changes in the MAP sampling design and its implications for the power analysis and the minimum detectable differences.

3.2 Experimental Design

Appropriate experimental design is critical to detecting change in PMs. While there are a number of texts on this topic (Fisher 1935, Kempthorne, 1952, Cochran and Cox 1957, Gilbert, R.O. 1987, Krebs 1989, Underwood 1997, Sit and Taylor 1998; and Zar 1999) they all recommend including: 1) a clear statement of the goals and objectives, factors, and factor levels to be investigated; 2) a description of the amount and type of replication; 3) the method of randomization, 4) the rationale for stratification and blocking; 5) hierarchal sampling at different temporal and/or spatial scales; and 6) a discussion of statistical inference.

Statistical inference is particularly important in the experimental design because it describes the set of assumptions that link the data collected on the sampled population to the target population of interest. The validity of statistical inference in an experimental setting should be considered carefully to assure that sufficient data pertaining to the objectives are collected and can be analyzed in the most efficient way (Deming 1953, Box et al. 1978, Hahn and Meeker 1993, and Sit and Taylor 1998).

3.3 Environmental Monitoring Design Criteria

Studies assessing ecological trends using long-term monitoring data (e.g., Olsen 1997, Dixon *et al.* 1998, Schmitt R.J. and Osenberg C.W. 1996, Olsen and Smith 1999, Manly 2000, Olsen *et al.* 1999) also address topics such as stratified and multi-stage sampling, repeated sampling from the same site or new sites each year; design-based and model-based inferences, assessment of status and trends, testing for serial correlation and causation, the role of power analysis, and detection of unforeseen responses (Phillipi 2003). A summary of recommended monitoring design criteria for inclusion in MAP assessments to guide experimental design are outlined here.

- A clear and unambiguous statement of goals and objectives that includes explicit statements of the target population of interest and the environmental variables to be monitored.
- A description of how inferences are to be made about the target population from the sampling design, what types of data will be used, and how representative the sampling sites are (see Section 5: Measuring Change from Reference Condition).
- It is crucial that the sampling design be able to accommodate change as objectives change over time. This is crucial if the results of the sampling design tell the investigator that the sampling design is insufficient to measure change or assess hypotheses. Additionally, management changes may occur over time necessitating the refinement of a sampling design. However, given the considerable spatial variation in the system, complete random sampling is unlikely to be the most powerful design to detect change for a given sampling effort. Stratification and other more general forms of multi-stage sampling are useful for leveraging ancillary data to provide more powerful estimates of variability, but random sampling within strata or sampling units is still required.
- If causation is important (e.g., linking the causal pathways from the conceptual ecological model), the monitoring design should be supplemented by experimental work to address specific hypotheses about causation (Olsen and Schreuder 1997, Olsen et al. 1999).
- It is crucial that the sampling design accommodate change as objectives change over time (Fuller 1999). However, designs that are too finely tuned while being the most powerful to detect specific a priori responses, greatly reduce the ability of the design to capture unforeseen responses (see Section 10: Future Uncertainties) and limit the ability to subsequently restructure the monitoring program in response to modified objectives or resource levels (Overton and Stehman 1996, Olsen and Schreuder 1997). These considerations are particularly important to successfully implementing the adaptive management strategy which is critical to CERP.

3.4 Estimating Variability

Several sources of variability common to monitoring data must be addressed as a prerequisite for determining whether a measured change in the performance criteria is different from its expected variability (i.e., baseline variability). Variability in all indicators should be characterized prior to detailed analyses to measure change. Measuring and characterizing variability are prerequisites for conducting many statistical analyses. However, in the absence of sufficient data to adequately characterize baseline, it will be incumbent upon the scientists to demonstrate that the observed changes are measurably different from baseline variability. As part of measuring variability, it will be important to develop strategies addressing, at the minimum, spatial and temporal (annual and interannual) variability and episodic events (e.g., tropical storms, hurricanes, etc.), that influence variability in South Florida systems.

3.4.1 Parametric and Non-parametric Techniques

When PIs provide data for use in the adaptive assessment process of the CERP, certain basic guidelines must be followed. These guidelines are necessary to ensure that the AT can carry out a timely evaluation of the assessments as well as track progress towards established restoration goals associated with CERP PMs and IG/IT.

A statistically robust monitoring and experimental design is necessary for any analyses. The PIs will be expected to determine the <u>distributional characteristics</u> of the data, providing a histogram plot and information to determine whether the data approximate a normal distribution. Statistical <u>tests for normality</u> (i.e., skewness and kurtosis) must be applied. If it is necessary to transform data, the PI will need to document that the transformation was successful in approximating a normal distribution prior to conducting parametric statistical tests (e.g., mean, standard deviation, standard error, etc.). Two examples are outlined in (Appendices B and C) for phosphorus levels in Lake Okeechobee and fish biomass estimates in Shark River. Both examples appear to meet the normality and homogeneity of variance assumptions.

However, parametric statistics are frequently inappropriate for ecological data, especially for rare, patchily-distributed animal species. Therefore, given that it is highly unlikely that environmental data collected in the MAP are going to be normally distributed, it is likely that non-parametric statistical tests would be an important component of the MAP. While transforming data may reduce variability and the influence of outliers, it does not guarantee normality. Non-parametric approaches, on the other hand, relax the normality assumption and are generally more robust than parametric methods. Regardless of the approach used, the PI is expected to provide a rationale for the selection of the analysis method. Additional resources addressing parametric versus non-parametric techniques are listed in Section 3.7 Additional Applicable Resources.

3.5 Power Analysis

The overall performance of the CERP monitoring program will be evaluated at periodic intervals. A useful tool for conducting this exercise is power analysis, which, depending on data properties and other factors, can be used to detect statistically significant differences in a performance measure. Statistical power can be defined as the probability of correctly detecting an effect (when it exists) or correctly rejecting a false null hypothesis, or avoiding type II error (see Appendix B), and depends on factors such as alpha level, magnitude of the effect, sample size, and sample variance. The appropriateness of the selected hypothesis for testing the ecological feature or process being examined and whether the data meet the assumptions of the selected statistical test are critical to the successful application of any statistic.

Statistical power analysis can be conducted before initiation of data collection programs, experiments, or management actions (*a priori* analysis), or following their completion (*a posteriori* analysis). Both approaches have advantages and limitations. *A priori* analyses, as advocated by most authors (e.g., Peterman 1990), are useful for making sampling intensity and allocation decisions. Limitations of *a priori* analyses stem from poor understanding of: (1) the response variable(s); (2) biologically significant effect sizes; and (3) extent of spatio-temporal variability. *A posteriori* or post-experiment power analyses are usually used to interpret a test that failed to detect an effect (i.e., a statistically non-significant result). This can result either from the *de facto* absence of an effect or from a low probability of detection.

Hoenig and Heisey (2001) argue that *a posteriori* analyses are useful for explaining observed data, but are not useful for data-analysis purposes and can result in seriously flawed interpretations of statistical analyses. They suggest that emphasis should be placed on confidence intervals, appropriate choices of null hypotheses, and equivalence testing.

Power calculations for a number of tests (e.g., t-tests, chi-square tests, ANOVA, ANCOVA, regression, correlation, multivariate methods) are available in the literature (see Appendix B). However, performing power analyses can be problematic when dealing with rare and/or patchily distributed animal abundances, which are typically dominated by zero values. This situation violates the normality assumption and forces the researcher to use a combination of statistical distributions (e.g., delta-log normal, delta-Poisson, etc.) to describe the data and estimate the composite variance of presence/absence and non-zero observations. Power analysis guidance for this common situation is not widely available.

The text that follows serves as guidance in situations where data generally meet (or can meet through transformation) the assumptions of homogeneity of variance and normality. It generally does not apply when nonparametric or multivariate methods are used. Refer to Noether (1987) for sample size determination for some common nonparametric tests.

Conventional power analysis uses the observed variance among replicates to estimate the minimum treatment effect detectable (difference from reference state). If the magnitude of the detectable difference among treatments is larger than the range of differences expected due to changes in Everglades hydrology/water quality, then revision of the sampling design should be considered to improve the statistical power of proposed analyses . Power analysis can be useful for determining whether data derived from the MAP sampling can be used to detect the desired changes in PMs. In addition, its inclusion in MAP may help determine if the sampling design is robust enough to detect the direction and magnitude of desired change within the desired timeframe. Methods for conducting power analyses can be found in standard statistic textbooks or in a wide range of computer programs and most recently on many university mathematics department web sites (Appendix B).

If variance estimates are not available, then an initial step in the MAP research process is to conduct a pilot study to obtain the information needed to estimate variance. The results from the pilot study will be used to measure and analyze the variability of the specific PM. This measure of variability will be used to conduct a power analysis to determine whether the current sampling design can detect the specified magnitude of change in the PM with the desired degree of confidence.

The utility of power analysis and its application in CERP is illustrated in the analysis of fish data collected in Shark River Slough. Trexler's 1997-2001 sampling data (Trexler, *et al.* 2001) provides an estimate of the across-year standard deviation of roughly 43 percent in raw biomass. Power analysis, conducted using this estimate of variance, shows the percent annual increase in biomass detectable at the 95 percent confidence level as a function of years of sampling.

Table 1. Minimum detectable change (a = .05) in biomass as a function of years of sampling. (After Phillipi 2003.

Years	% Change Detectable
2	171
3	65
4	37
5	25
9	9.5
10	8.1
13	5.4
14	4.8
19	3.0
25	2.0

Figure 3-1: Minimum detectable change in biomass as a function of years of sampling (Phillipi 2003)

This type of analysis is important to understanding the ability of RECOVER to detect quantitative changes in PMs given the current MAP design. For example (Figure 3-1), after two years of sampling using Trexler's experimental design you would only be able to detect a 171 percent change in biomass with 95 percent confidence. Any change less than that would not be considered statistically significant. Similarly, it will require between 9-10 years of monitoring data to be able to detect a 10 percent annual increase in biomass to exclude no trend, 14 years are required for detecting a 5 percent annual increase, and 25 years are required for detecting a trend of 2 percent increase per year in biomass.

However, it is incumbent upon the PI to determine what degree of measurable change is deemed of ecological significance and not just statistical significance. For example, if a population has a large (>25 percent) intrinsic interannual variability the sampling design need not be so rigorous that it requires a detection level at a 10 percent change with 95 percent confidence. Rather, a sampling design that detects 25 percent change with 90 percent confidence might be sufficient given the wide interannual variability. The detection limits and confidence parameters will be expected to vary with individual performance measures.

Andersen (1998) discusses the importance of "errors of inference" relative to drawing erroneous conclusions from the analysis of management experiments and monitoring programs (Appendix B-3). Specific attention is given to Type I and II errors (statistical power analysis) and their importance in AM. A P-value of 0.05 is common practice in science, but it makes change very difficult to detect in highly variable data such as that collected by the MAP. Walker (2000), in his paper on total phosphorus increases to Everglades National Park, uses a P value of 0.10. Currently, the compliance regime for the Park is based on p-value of 0.10. Therefore, a higher P-value (> 0.05) could be considered for routine change detection using field data.

Anderson (1998) provides the following recommendations to scientists and managers for addressing the issue: recognize that errors of inference are unavoidable, but their frequency can be controlled by experimental design; include *a priori* power analysis in all monitoring designs; realize that non-significant results should report the effect of size and power of the design; use *a posteriori* power analysis also; where current experimental designs lack power, replace them with new, more powerful methods (e.g. BACI (Before-After-Control-Impact) - paired designs, Underwood 1994); conduct pilot studies to improve power of large experiments; and finally, because not enough is known about potential response variables (e.g., biologically significant effect sizes and spatial and temporal variability), it is important to conduct long-term monitoring for important variables.

Power analyses are strongly recommended, but their results should not be the only criteria for adoption or rejection of the ability to measure change in a given parameter of interest (i.e., PM). Rather, researchers should design their programs around a given parameter for its ecological relevancy to CERP-related changes. For many of ecological components, responses to CERP activities may well emerge in "fits and starts". Alternatively, some of the best indicators of CERP performance may be highly variable now (i.e., pre-CERP) because they are under stress, but not necessarily later (i.e., post-CERP) if stress is reduced. Power analysis can be an extremely useful tool for designing or modifying a monitoring program, but due caution is warranted before abandoning a given PM based on power analysis results alone. Monitoring programs should strive to collect as much CERP-relevant data as is practical and feasible, and researchers should be mindful that our current understanding of the system and the responses of its components is incomplete.

3.6 Additional Applicable Resources

- South Florida Wading Bird Report – September 2005 (Mark I. Cook & Erynn M. Call, Editors)

- Phillipi, T. 2003. CERP monitoring and assessment plan: stratified random sampling plan. SFWMD Agreement C-C20304A September 24, 2003.
- Society of Landscape Ecology

4.0 ESTABLISHING REFERENCE CONDITION

4.1 Purpose

The establishment of reference conditions prior to implementing CERP project activities is one of the primary goals of the assessment and it is against these reference conditions that changes in the PMs that occur during and following CERP projects will be evaluated. The reference condition, or baseline, will include a minimum of four cumulative years of MAP data supplemented with relevant non-MAP data covering longer time series. In addition, the reference condition must be cumulative in order to include the full range of variability including capturing the extremes. The issue is confounded because of the heterogeneity in the temporal and spatial expression of the different PMs. For example, because hydrology and water quality changes may occur rapidly, changes in these parameters relative to a reference condition can be ambiguous. Changes in ecological responses are generally slower, thereby complicating the establishment of reference conditions as well as detecting change. It is critical that data sets used to establish the reference conditions meet the MAP guidance criteria.

The objectives of this component of the guidance process follow.

1) Describe the non-MAP data sources used in the assessment
2) Ensure that monitoring data meet the guidance criteria or provide rationale that justify their inclusion
3) Ensure that the data have been entered into the CERP's centralized environmental system, and updated appropriately.

4.2 Historic and Current Databases

The availability and quality of the monitoring data used to characterize the CERP reference conditions (baseline variability) for each of the performance measures may vary and may include MAP and non-MAP data (e.g. project level monitoring data). PIs and Module Groups should consider including all appropriate non-MAP data in their assessment efforts and consequently, identification and selection of the monitoring data used for establishing reference conditions and measuring changes in PMs must meet minimum criteria.

4.3 Criteria for the Use of Non-MAP Databases

To ensure that a wide range of data, in addition to those collected by the MAP, is available to estimate reference conditions for the PMs, we propose the following criteria be used to assure the data are comparable to, and consistent with, those being collected by the MAP. Observational data (existing water quality monitoring in the Greater Everglades), should include one or more of the following bullets listed below:

- The data were collected for sufficient duration to provide estimates of both intra- and inter-annual variability.