6.0 EVALUATING HYPOTHESES AND INTERIM GOALS

6.1 Purpose

The guidance in the previous sections has focused primarily on detecting and measuring the magnitude and direction of change in individual performance measures and to determine if the changes are consistent with expected responses described in the CERP hypotheses. The focus of this section is not on the individual performance measures, but on the MAP hypotheses, which explicitly describe the causal pathways (i.e., stressor-response relationships) that explain how the system at the Module level behaves. Since there are multiple performance measures supporting the module-level hypotheses, this section will present a variety of approaches for integrating and interpreting physical, chemical, and ecological data (i.e., multiple performance measures) to assess the causal inferences comprising each hypothesis. If the trends or weight-of-evidence assessments of performance measures or hypotheses do not correspond to predicted responses, the Module Group will be expected to provide a scientific explanation. The assessment of module-level hypotheses will then be used to evaluate progress toward achieving IG/IT. The following sections will present the various methods and models that can be applied to assessing module-level hypotheses.

6.2 Approaches

The MAP is structured around the use of CEMs to develop working hypotheses clusters that describe how various defining ecological components of the system have and will continue to respond to stressors. One or more PMs developed for the critical stressors and defining ecological characteristics are then used to track the progress of the restoration in meeting its goals. The challenge for scientists will be to integrate and interpret the multiple PMs that comprise the hypotheses for that particular spatial domain.

At the module and system level of synthesis and integration, a range of analysis tools will be employed individually and in combination. These tools include, but are not limited to, CEMs, statistical regression models, numerical simulation models, and lines-of-evidence (LOE) and weight-of-evidence (WOE) analyses. The choice of analysis methods will depend on the nature of the performance measures, the hypotheses, the types and quantity of data, and the management goals and context. Thus, a suite of analysis tools will be applied to answer specific questions regarding the performance of various components of the system. The challenge is how to best synthesize and interpret the results of these multiple tools to provide an overall assessment for the Module or Total System.

6.3 Statistical Analysis

A variety of statistical tests and quantitative approaches have been used for investigating possible relationships between independent (stressor) and dependent (effects) variables acquired from monitoring programs. The basic types of statistical methods that can be used to examine the types of relationships that comprise the hypotheses in the CERP CEMs include:

- Simple descriptive statistics (i.e., means, ranges, variances, etc.)
- Exploratory statistics (i.e., multivariate correlation analysis and logistic regression)
- Statistical modeling (i.e., stressor-response relationships)
- Goodness-of-fit or maximum likelihood methods
- Trend analysis.

Patterns and trends in monitoring data can be estimated using parametric and nonparametric statistical tests. Typically a parametric test (e.g., single and multiple linear and non-linear regression, etc.) involves the regression of transformed and non-transformed data with time. Non-parametric tests use non-transformed data.

However, while statistical analysis can reveal relationships among data and provide insight to possible underlying mechanisms, establishing causation requires integration of additional knowledge. Because stressors generally co-occur and co-vary with each other and with natural environmental attributes, relationships between a candidate cause and a biological variable may be due to factors other than the candidate stressor. Statistical hypothesis testing was designed for analyzing data from experiments where treatments are independent, replicated, and randomly assigned. The application of these statistical tests to data from observational field studies, where treatments are: seldom replicated; randomly assigned to experimental units; at best pseudo-replicated; must be viewed with caution.

An overview of parametric tests, non-parametric tests, and the use of censored data are presented in the following section as a "primer" for approaching CERP analyses.

6.3.1 Parametric Tests

Use of statistical correlations, multivariate and logistic regression, and trend analysis are the most common statistical approaches for attempting to identify associations among environmental variables particularly in monitoring studies (Stowe *et al.* 1998, Johnson and Collier 2002). One of the most useful techniques is logistic regression analysis that is commonly applied to binomial or proportional data. A major advantage of logistic and multivariate regression analysis is that they allow the significance of relationships between the stressors and biological effects to be determined while adjusting for possible other influential treatment factors (e.g., season, temperature, etc.).

In time-series analyses, a linear trend is estimated by regressing a response variable (i.e., flow) as a function of an explanatory variable (i.e., time). In most cases, data must be log-transformed in order to improve linearity. If this transformation is used and assumptions of normality, constant variance, and independence are met, then a null hypothesis of zero slope over time can be tested. If the slope is significantly greater than zero, the null hypothesis is rejected and one can conclude that a linear trend over time has occurred. Because regression models do not account for error or dependence in the estimates, a conservative p-value of 0.01 or less is recommended for determining significance.

For purposes of CERP, a trend may be defined as a change over time in the mean of the parameter of interest. Trends are identified when the magnitude and direction of change in a parameter of interest can be clearly differentiated from natural background variability. Historic datasets or monitoring programs therefore must be of sufficient duration and spatial extent to quantify the range of natural background variability. The length of the time series required to identify a trend will depend on the properties of the ecosystem component under study.

For example, the time-series required to define a trend in fast-response variables (e.g., water quality parameters downstream from a structure) will be shorter than the time-series required to determine trends in slow-response variables (e.g., sediment elevations, biological response, etc.). Many of the changes in CERP will generally occur over long-time scales (e.g., multi-year, decadal). Monitoring programs designed to measure trends in these ecosystem components must be of sufficient duration in order to differentiate the effects of CERP from interannual or inter-decadal variability.

6.3.2 Non-Parametric Tests

The following non-parametric tests are specifically designed to address tends in seasonal data that are particularly relevant to the wet-dry season patterns characteristic of South Florida. The Seasonal Kendall (SK) test (Hirsh 1982 1992) is a non-parametric test for seasonal trends for a mono-tonic linear trend that is resistant to outliers and is not dependent on the normality of the data set. By comparing only the data from similar seasons, the test also reduces seasonal effects when testing for trends

The Kendall-Theil (KT) analysis is another non-parametric test for monotonic linear trend by use of pairwise comparison and a Kendall's tau test for significance, similar to the Seasonal Kendall test. This test accounts for seasonality by calculating pairwise slopes on data within the same season with the overall trend slope computed from the median of the seasonal slopes. The KT analysis differs from the SK in that the KT seasonal adjustment method allows for increased power in the slope estimate when compared with the SK test.

6.3.3 Censored Data

A large number of data values that are below the detection limit (i.e., censored data) can adversely affect the estimation of the slope of a trend by not allowing for corrections because of variations in the variable of interest. The following guidance developed by USGS may be useful when dealing with censored data.

- <5 percent censored data – censored values will be assigned one half of the detection limit. For all trend tests, it is recommended that the p-value, the slope estimate, and the magnitude (percent change over time) will be reported.

- Between 5-20 percent censored data – three separate trend analysis tests of the data will be performed. First, on the raw data file with censored data set to half the detection limit; second, all censored data set to 0 (KT only); and third, all censored data set to the detection limit (KT only). In these instances, it is recommended that the highest p-value, the range in slope estimate, and a range in the magnitude of the trend be reported.
- >20 percent censored data – trend results will not be reported.

6.4 Simulation Modeling

Simulation models are basically mathematical models that attempt to integrate and show the interactions between multiple environmental components (i.e., physical, chemical, and biological) of the ecosystem and are becoming increasingly important in environmental management. However, there is an ongoing debate as to the relative merits of simple models that have been statistically fit to data versus more complex models that attempt to simulate chemical, physical, and ecological processes (Shipley and Peters 1991; Reckhow and Chapra 1999). There are well-founded arguments in support of each approach. If the relevant processes are understood and can be expressed mathematically, then a process-based model should yield reliable assessments and forecasts of system behavior. However, if not, then this approach becomes an exercise in calibrating an over-parameterized model to limited data rather than a true mathematical representation of our process knowledge. Statistical methods, on the other hand, are rooted in empirical observations that rarely have an explicit mechanistic basis. Consequently, no single model is likely to satisfy all needs. The following is a brief description of some of the modeling approaches available to CERP with examples of their applications.

At the module and system level of synthesis and integration, a range of modeling tools, individually and in combination, will have to be employed (e.g., CEMs, statistical regression models, and numerical simulation models) depending on the performance measures, hypotheses, types and quantity of data, and management implications of interest. The challenge is how to best synthesize and interpret the results of these multiple outputs to provide an overall assessment for the Module or System as a whole.

6.4.1 Modeling Approaches

The recovery of the Everglades will be the result of a number of interacting processes operating at multiple spatial and temporal scales. Modeling approaches, representing varying degrees of complexity, data requirements, and spatial scales may be used to assess changes in and the integration of multiple PMs. These include small spatial scale statistical models and habitat suitability index models (HSI), intermediate- scale models (e.g., Adaptive Ecosystem Assessment Models, Bayesian Network Models, Simple Ecological and Social Dynamics Models) and large-scale models driven by extensive data requirements [e.g., Lake Okeechobee Water Quality and Hydrology Models, Across Trophic Level Systems Simulation (ATLSS) and Everglades Landscape Models (ELM)].

However, the individual models developed to appropriately represent each of these scales and processes are not easily combined into a single assessment of evaluation model.

Intermediate scale modeling (as opposed to intermediate scale models) involves constructing a wide variety of models for assessment, parsing out the needed information from each, and then integrating the results. This approach must serve both as a assessment (post-CERP project implementation) and an evaluation tool (prior to project implementation). Evaluation (forecasting) is also done as part of the adaptive management process. The foundation of both evaluation and assessment rests on the ability to predict future results of project implementation while including estimates of uncertainty. The focus of the this Gudiance is to provide the assessment framework (e.g., data, statistical methods and models) that are appropriate for both evaluation and assessment.

Walters *et al.* (1992) demonstrated the use of his Adaptive Environmental Assessment Model (AEAMs) as a tool for screening water policy alternatives for ecological restoration in the Everglades. Carpenter *et al.* (1999), has developed simple ecosystem management integrated simulation models (Ecological and Social Dynamics Models) to explore and illustrate interactive dynamics of socio-ecological systems in a lake subject to phosphorus pollution. The Bayesian network modeling approach employs a graphical structure, analogous to the CEMs, to explicitly represent the variables and causal relationships involved in the relevant processes (Borsuk *et al. in press*). In so doing it provides a framework to integrate a variety of models representing a number of interacting processes operating at multiple spatial and temporal scales. This graphical approach explicitly represents cause-and-effect assumptions between system variables that may be obscured under other approaches. Brief discussions of these types of modeling approaches can be found in the attached Appendix C.

It is unlikely a single model will capture the explicit dynamics of parameters and processes occurring over multiple temporal and spatial scales. However, simulation models should be used as one of a spectrum of quantitative tools in the assessment and evaluation process that can be use to estimate and predict changes in multiple sets of variables. A suite of modeling tools, individually and in combination, may be employed (e.g., CEMs, statistical regression models, and numerical simulation models) depending on the performance measures, hypotheses, types and quantity of data, and management implications.

6.4.1.1 Bayesian Network Models

The basics of Bayesian networks and their application to environmental prediction are well described in the literature (Reckhow 1999, Borsuk *et al.* 2002). The Bayesian network modeling approach employs a graphical structure, analogous to the conceptual ecological models employed in CERP (e.g., Wetlands [2005] Vol. 25, No. 4), to explicitly represent the variables and causal relationships involved in the relevant processes (Borsuk *et al. in press*). In so doing it provides a framework to integrate a variety of models representing a number of interacting processes operating at multiple spatial and temporal scales. This graphical approach explicitly represents cause-and-effect assumptions between system variables that may be obscured under other approaches. Note: hypothesis cluster concept for inferring causal relationships

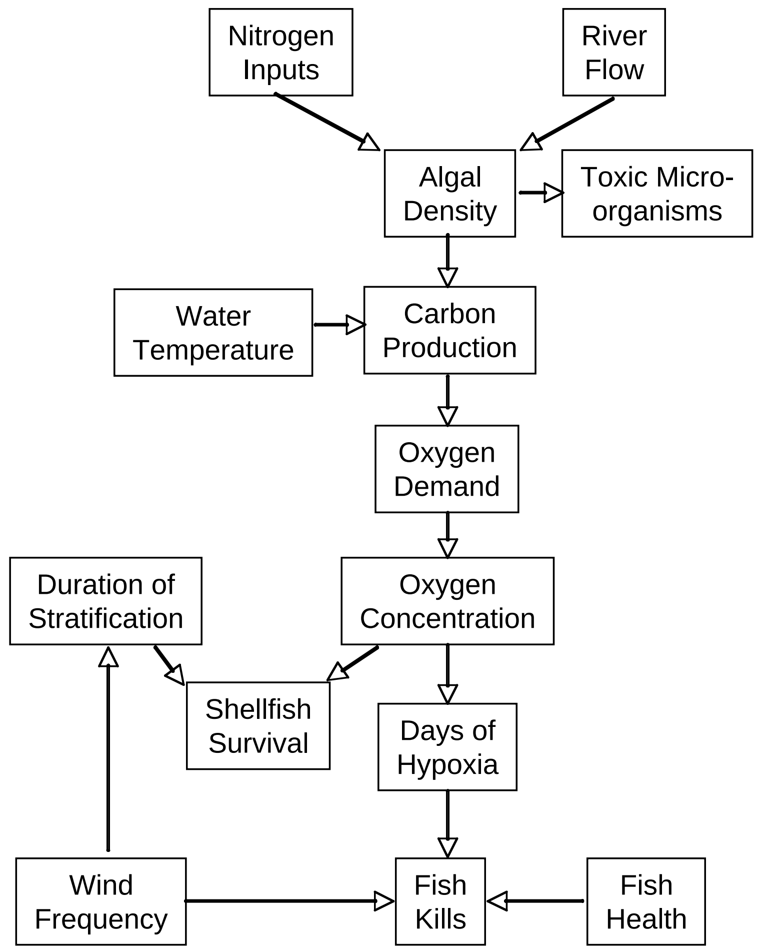

Figure 6-1. Causal network of important factors involved in the eutrophication of the Neuse estuary.

Figure 6-1: Causal network of important factors involved in the eutrophication of the Neuse Estuary

Briefly, a Bayesian network is a graphical depiction of the relationship among the most important variables in the system of interest (Figure 6-1). In this illustration, the variables are represented by square nodes, and dependencies between one variable and another are represented by arrows. The conditional independence, implied by the *absence* of any connecting arrows, greatly simplifies the modeling process by allowing separate sub-models to be developed for each conditional relationship indicated by the *presence* of an arrow. These sub-models may be derived from any combination of process knowledge, statistical correlations, or expert judgment, depending on the extent of information available about that particular relationship.

In a Bayesian network, each dependency indicated by an arrow is characterized by a conditional probability distribution that describes the relative likelihood of each value of the down-arrow node, conditional on every possible combination of values of its parents. A node that has no incoming arrows is said to have no parents, and such a variable can be described probabilistically by a marginal (or unconditional) probability distribution. A useful practice when developing conditional distributions in the network is to view them as functional relationships among variables [Pearl 1999].

Because the Neuse Estuarine system was studied at multiple scales, a modeling framework was necessary that represented all relevant system processes (Borsuck *et al.* 2004). The Bayesian network analysis framework was chosen in this study because of its ability to integrate sub-models of disparate scales. Choosing the various scales of representation in a Bayesian network should be a dynamic and iterative process. This is because the scales which are chosen to represent key features of the ecosystem are often imposed by observational capabilities, technological, or organizational constraints [Levin 1992] which may evolve over time. Further, the scale of prediction should correspond to the needs of decision-makers, which may also change with time as they gain understanding of the problem.

6.4.2 Model Applications

There are a number of examples of model applications under development or being used in South Florida. Below, three (one from the Greater Everglades, another from Lake Okeechobee, and one from Florida Bay) are presented here that encompass the spectrum of intermediate-scale modeling tools considered for use in modeling ecosystem responses to CERP implementation in South Florida.

6.4.2.1 Greater Everglades Wading Bird Model

The CEM (See Section 9.2.1) provides a description of the causal linkages between hydrological patterns, the abundance and distribution of prey species and the success of wood stork and spoonbill nesting. An example of how the CEM can be used to develop an intermediate-scale modeling framework is illustrated in Figure 6-2. The proposed model for wading birds presents a different way of representing interactions by showing not only the hydrology model, which acts as the foundation, but illustrates the various sub-models that spinoff and quantify the functional relations illustrated by the conceptual model sub-hypotheses for vegetation, fish, and wading birds. The model can be used iteratively by applying five to ten years of data to the previous five to ten years worth of data, allowing scientists and managers to compare model predictions with the annual assessments of the monitoring data. The ability to iteratively refine the model strengthens it as a predictive tool.

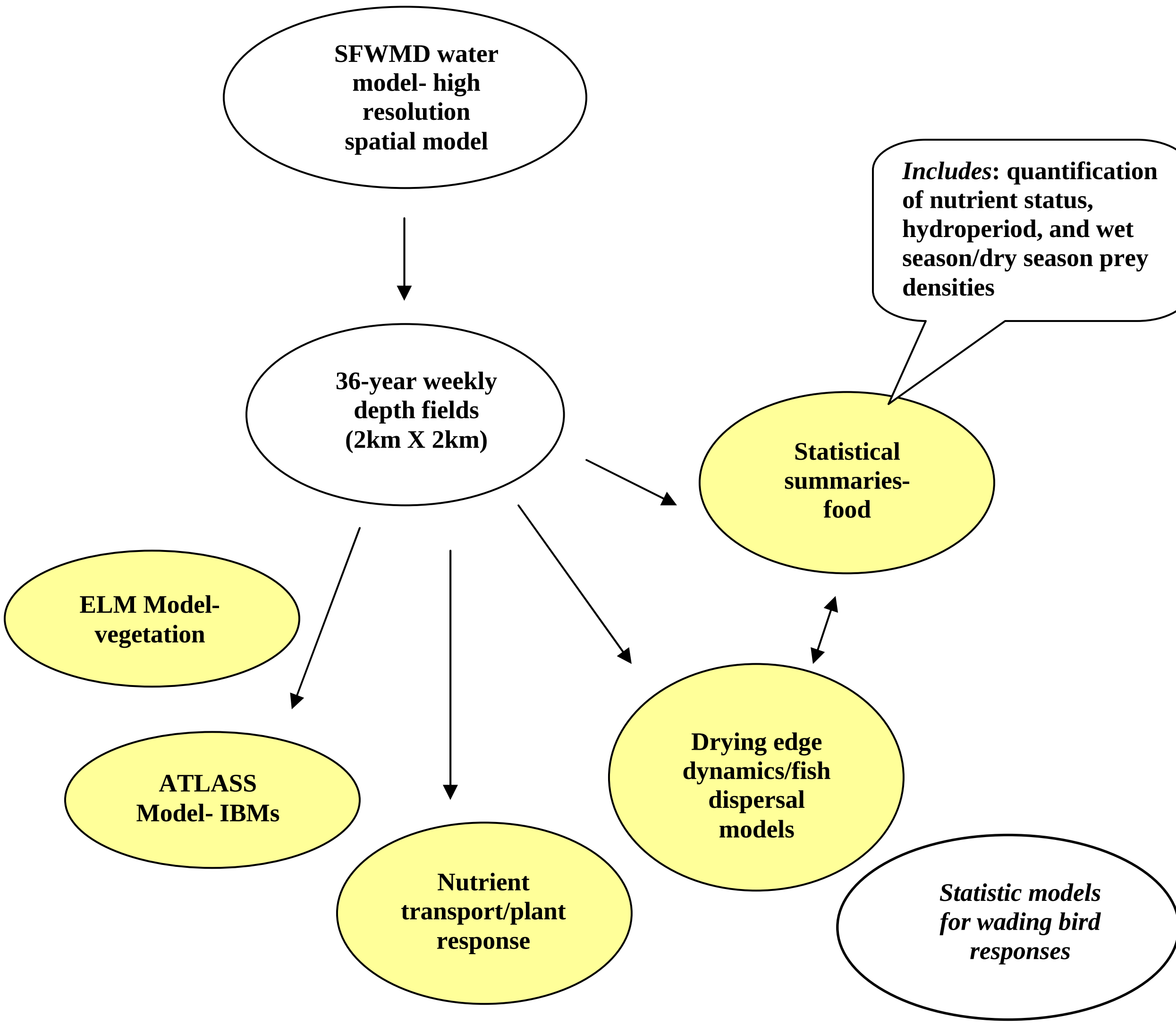

Figure 6-2: Hydrology Effects of CERP on Wading Birds and Their Prey Base

6.4.2.2 Lake Okeechobee (Hypothesis Assessment Model)

A second example can be used to illustrate the testing of a hypothesis from the Lake Okeechobee MAP module: "At lower lake stages, increased light penetration results in an increase in the biomass and spatial extent of submerged aquatic vegetation, and in turn, increased recruitment success of largemouth bass in Lake Okeechobee" (RECOVER 2004). The first component to be analyzed is the relationship between stage and plant biomass. A simple multiple regression model is used to predict plant biomass from measured water depth and suspended solids (Havens *et al.* 2004). In addition, a STELLA[PE1] Model developed for submerged plants in the lake tracks the plant biomass changes as a function of lake stage (e.g., water depth). The third component compares submerged vegetation with largemouth bass size structure. The model predicts successful bass recruitment when biomass of submerged vascular plants becomes high,, and very low recruitment of bass when submerged vascular plant biomass is low (no plants, or dominance by *Chara spp.*) (K. Havens, personal communication).

6.4.2.3 American Crocodile (Evaluation Model)

The American crocodile is an endangered species in the Everglades ecosystem. A key component of the CERP is re-establishing freshwater flows to mangrove estuaries that are critical crocodile habitat (Chartier *et al.* 2005). The purpose of this study is to forecast the crocodilian population responses to alternatives for restoration using integrated hydrological and ecological models. A salinity-based habitat model was developed to evaluate restoration of freshwater flows to the mangrove estuaries and the Southern and Inland Coastal Systems Model (SICS) was used to generate alternative freshwater flows, depths, and salinity for the ecological habitat model. The habitat model is comprised of three processes: hatchling growth and survival, forage production, and forage availability (Chartier *et al.* 2005). The SICS and habitat models were used to investigate: 1) how does the amount of freshwater entering Florida Bay affect suitability of crocodile habitat; and 2) what would be the impact of restoring the historical water flow into the ecosystem? By using the coupled intermediate scale models, the authors concluded that the relative suitability of crocodile habitat in northeastern Florida Bay is driven by the amount of freshwater flow to the Bay. Second, the location and delivery of water may be less important than the amount of freshwater flow. This study illustrates the concept of integrating statistical relationships between salinity and flow forecasts from a hydrodynamic, physically based simulation model and a habitat suitability index to determine the areal extent of suitable habitat for the crocodile (Chartier *et al.* 2005).

6.5 Causal Inference: Lines-of-Evidence Approach

The focus of the MAP assessments is to evaluate hypotheses that describe the functional relationship between one or more stressors and ecological attributes of interest Implicit in the MAP hypotheses is the concept of causation. While it is often difficult to establish causation definitively from environmental data, one can develop criteria for supporting or rejecting the hypothesis using weight-of-evidence or multiple-lines-of-evidence strategies.

Lines-of-Evidence (LOE) (sometimes referred to as Weight-of-Evidence [WOE]) approaches have a long history in courts of law. While criminal law in the United States refers to proof beyond a reasonable doubt, civil law refers to "preponderance of evidence". Both criminal and civil actions depend on a WOE approach (Chapman *et al.* 2002). Formalized use of WOE frameworks for integrating and interpreting multiple lines of evidence in environmental sciences is relatively recent and includes individual LOE as well as combined LOE. Hill (1965) suggested nine criteria for inferring causality: strength of association, consistency, specificity, temporality, biological gradient, plausibility, coherence, experimental evidence, and analogy. Mosteller and Tukey (1977) noted three criteria: consistency, responsiveness, and mechanism that must be satisfied to infer cause-effect relationships. Table 6-1 summarizes different lines of evidence for inferring potential causal relationships.

Table 6-1: Comparison of criteria for establishing causality from different sources

Fox (1991); Suter *et al*. (2000); Burton *et al*. (2002)	**US EPA (2002)**	**Suter *et al*. (2002)**	**Hill (1965)**
Strength of association (between stressors(s) and effects (s))	Effect magnitude (strength of link)	Complete exposure pathway	Strength of stressor effects association
Consistency of association (between stressor(s) and effects(s))	Consistency of association (at multiple sites)	Consistency of association	Consistency of response in different circumstances
Gradient (spatial or temporal concentration or dose-response relationship.	Co-occurrence (spatial/temporal correlation)	Co-occurrence Temporality Biological gradient	Temporality – stress precedes response Biological gradient
Plausibility (probable cause-effect mechanism)	Experimental confirmation (field or lab)	Plausibility Experimental	Plausibility – stress response relationship is consistent with theory
Specificity (association specific between stressor(s) and effects(s))	Specificity (stressor causes unique effect)	Specificity of cause Predictive performance	Specificity – uniqueness (diagnostic) of stressor-effect relationship
Analogy (other cases with same relationship of stressor(s) to effects(s)	Plausibility (likelihood of stressor-effect linkage)	Analogy	Analogy – similar stressors associated with similar effects
These cells intentionally left blank			Coherence – stressor effects hypothesis consistent with available evidence Experimental evidence supports stressor-effect causal relationship

Often these criteria are weighted equally when assessing causality. However, Menzie *et al.* (1996) used a weighted approach in an ecological risk assessment (ERA) context. Here WOE is defined as a weighted multiple-lines-of-evidence for assessing the strength of the relationship by which measurement endpoints (e.g., PMs) are related to assessment endpoints (e.g., valued ecological attributes, defining ecological characteristics) based on weight, magnitude and occurrence. Ten separate judging attributes are used, which may be equal or weighted: degree of association; stressor/response; utility of measure; data quality; site specificity; sensitivity; spatial and temporal representativness; quantitative measure, and standard methods of measure. The Menzie *et al.* (1996) approach also allows for varying degrees of best professional judgment, which allows the quantitative matrix approach for weighting attributes to be replaced by a qualitative weighting system. This approach might be particularly applicable to assessing the importance of multiple PMs for evaluating specific hypotheses clusters when data is not sufficient for a quantitative assessment.

Approaches for combining lines of evidence can be broadly divided into five different categories: indices, statistical summarization, scoring systems, logic systems, and best professional judgment (Burton *et al.* 2002, Chapman 1996). Similarly, a variety of analytical methods can be used for analyzing each line of evidence including: trend analysis using regression analysis; ANOVA methods including BACI, covariate analysis, and multivariate regression analysis.

6.5.1 Decision Frameworks Based on Weight-of-Evidence (WOE) Methods

The linkage of multiple variables within a hypothesis (LOEs) to inform a decision is a multi-step process that involves weighing the strength of evidence that supports each relationship. This strength of relationship is established using criteria outlined in Koch's postulates (Suter 1993) that were originally developed to infer causality between pathogens and the manifestation of disease. These postulates are:

- The adverse effect must be regularly associated with exposure to the stressor
- The stressor must be found to co-occur in space and time with the receptor
- The adverse effect must be manifest in unimpaired species (i.e. healthy) following exposure to the stressor
- The stressor must be found in the affected species

Koch's causal postulates can be satisfied by applying the causal criteria described in Table 6-1. After these evaluations are conducted, individual expert qualitative judgments interpreting the relationships (i.e., potential causality) can be integrated using a tabular decision matrix. The results of these expert judgments can be summarized in a tabular decision matrix. For example, converting to ranks (e.g., 1 to 4 or "+" and "-" values) per Chapman (1990 1996) Grapentine *et al.* (2002), USEPA (2000).

There are a variety of qualitative and quantitative approaches for integrating multiple LOE into a weight of evidence based decision (Burton *et al.* 2002). The following table summarizes these approaches and discusses the strengths and limitations of each approach.

Table 6-2: Advantages and Limitations of Different WOE Approaches for Combining LOEs					
	Robustness	**Method**	**Sensitivity**	**Appropriateness**	**Transparency**
Qualitative	Low	High	Low	High	Low
Expert Ranking	Low	Moderate	Moderate	High	Low
Consensus Ranking	Low	Moderate	Moderate	High	High
Semi-quantitative ranking	Low	Moderate	Moderate	High	Low
Broad Scale Multiple WOE	Moderate	Low	Moderate	High	Moderate
Quantitative Likelihood	Moderate	Low	High	High	Moderate
Tabular Decision Matrices	Moderate	Moderate	High	High	High

All WOE approaches depend on data quality, study design, and expertise used in a manner to best infer the causal links between stressors and effects. The advantages and limitations of each approach can be generally characterized using the evaluation criteria in Table 6-2 where:

- **Robustness** = consistency in interpretation and decision-making irrespective of when and where conducted
- **Method** = ease of use, if the approach is to complex or expensive it is unlikely to be widely used
- **Sensitivity** = providing good discrimination not only between extremes but also for intermediate situations
- **Appropriateness** = broadly applicable, not restricted to specific conditions of environments
- **Transparency** = readily understandable at both the LOE and WOE levels.

Most of the literature on WOE or multiple LOE approaches simply "lump" the various LOE results in a non-quantitative manner. The effectiveness/accuracy of any LOE/WOE approach is contingent on: 1) the quality of the data; 2) the quality of the study design; 3) expertise of the principal investigator(s); 4) the severity of the disturbance/impact; and 5) a matching of the objectives to the data. Accurate WOE conclusions require excellence in factors 1, 2, and 3. In general, qualitative approaches for combining multiple LOEs can be categorized as having low robustness, high ease of use, low sensitivity, high appropriateness and application, and low transparency (Table 6-2).

A variety of expert ranking methods have been proposed (e.g., Swanson and Socha 1997, Calabrese *et al.* 1997, Bombadier and Bermingham 1999) that all rely on a panel of experts using Best-Professional-Judgment (BPJ) to arrive at an integrated assessment. However, the expert ranking system is not very robust, moderately difficult to use, moderately sensitive, highly appropriate/applicable, and not readily transparent.

The consensus ranking approach derives from the desire to include stakeholders in the decision-making process. The most common method to inform stakeholders is to use ranking methods displayed as tabular decision matrices (Chapman 1990,Chapman *et al.* 1996, and Menzie *et al.* 1996). Because of both variability and acceptability factors, ranking has evolved into a consensus-based process where stakeholders agree to the study design and how the data will be interpreted. Menzie *et al.* (1996) describes a consensus-based process for WOE that has both weighting and scaling factors that are agreed upon, a priori, by mangers, stakeholders, and scientists. The primary advantage of the consensus-based metric ranking approach is that it can achieve, *a priori,* a general consensus between stakeholders on the study design and interpretation in a WOE approach.

Semi-quantitative ranking approaches have been widely used both within LOE (e.g., benthic and fish metrics, etc.) and between LOE (physical, chemical, and toxicological data) for sediment toxicity. Other semi-quantitative approaches permit the adjustment of metrics for certain non-stressor covariates, such as stream size, in the computation of the Index of Biotic Integrity (Karr 1981). This approach varies widely in its design; however, in general it is not very robust, has moderate levels of methodology and sensitivity, while the appropriateness/applicability is high but the transparency is moderate to low.

Broad-Scale WOE incorporates multiple WOE approaches (e.g., qualitative, ranking, or semi-quantitative) to varying degrees. A variety of methods can be incorporated under this rubric that demonstrate how knowledge gained through time from a variety of case studies with similar stressors and comprised of multiple LOE may be used to establish causality and link environmental components in a decision-making framework (Lowell *et al.* 2000, Culp *et al.* 2000). The US EPA (2002) and Sture *et al.* (2002) describe a non-quantitative stressor identification evaluation process that sets out a framework for establishing the basis for impact using a WOE process. In this process, BPJ is used to score the several causal criteria (Table 6-1) using ranks (e.g., 1 to 4; or "+" or "-") but does not permit summing because that would imply equal weighting. This type of approach is moderately robust with methods being somewhat complex. Its sensitivity is moderate, however its appropriateness and application are high and the level of transparency is low.

Quantitative likelihood takes advantage of the fact that many LOE involve quantitative measurements. While this is often viewed as an advantage over qualitative information, the issue of interpreting the ecological importance of statistical significance must be addressed (Goodman 1999). The details of this approach are beyond the scope of this guidance but can be found in Burton *et al.* (2002) and cited references.

Tabular decision matrices for WOE were first proposed by Chapman (1990) and continue to be developed and refined (Chapman *et al.* 1996, Grapentine *et al.* 2002, Menzie *et al.* 1996). The advantage of tabular decision matrices is to provide information on individual LOE relative to, at the simplest level, a binary classification (e.g., altered or not). The tabular ranking approach has many advantages over the other ranking approaches, being moderately robust, has moderate methodology, and high degrees of sensitivity, appropriateness/applicability and transparency.

6.5.2 Application of Integration Approaches

For the purposes of illustration (Figure 6-3) we have selected one of the conceptual model MAP hypotheses from the Greater Everglades Wetlands Module that links MAP field data on hydrology, aquatic prey fauna, and wading birds. The measurable change, between pre- and post-CERP conditions is then estimated for each of the physical, chemical, and biological variable and/or their metrics using either statistical or modeling approaches. The outcomes from these analyses are referred to as lines-of-evidence (LOE) which are then combined into a weight-of-evidence (WOE) decision.

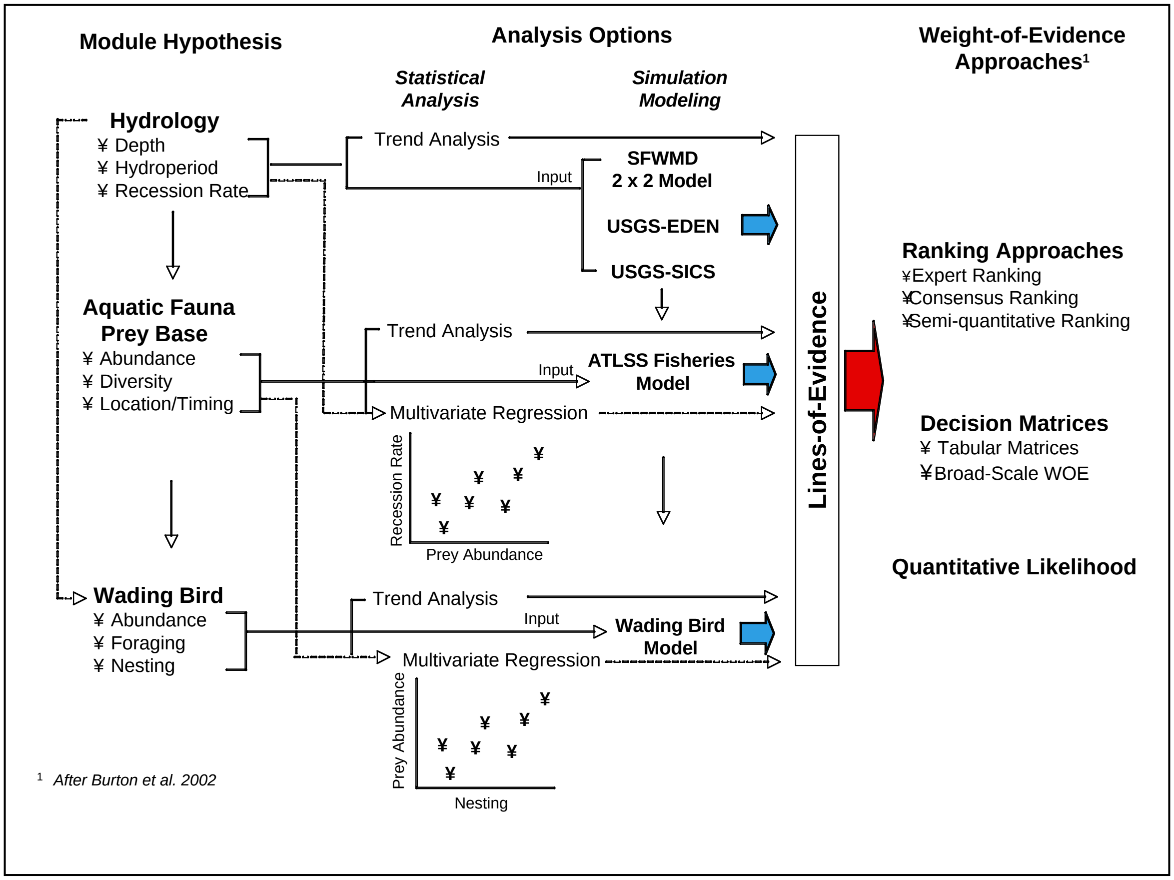

Figure 6-3: LOEs and WOEs for the Greater Everglades

In this example, all the attributes (physical, chemical, and biological) can be analyzed using a variety of parametric and non-parametric statistical approaches as discussed above. At the very least, each variable, nine in this example, will be examined over time to illustrate the trends for pre- and post-CERP time. In addition, multiple variables can be analyzed using multivariate statistical methods to examine relationships between stressor and response variables (e.g., as many as twelve in this example, as well as their important non-stressor covariates to understand and interpret the status of the hypotheses.

Each of the nine variables can be used as input terms to various modeling modules that can be combined to address a specific hypothesis. In addition, the output from the simulation modeling is considered a LOE and will be available for interpretation. Each of these analyses can be treated as a line-of-evidence (LOE) and combined into a weight-of-evidence (WOE) assessment to determine the status of the hypothesis or interim goals.

In the example illustrated in Figure 6-3, recession rate was chosen as a hydrological variable because of its importance in controlling the density and spatial distribution of aquatic prey species. Information on this variable, generated from field data or modeling, can be used in three ways: 1) it can be analyzed as a trend (pre- and post-CERP) using a BACI or similar approach; 2) as input data into an ecological model; and 3) as the independent variable in a multivariate regression analysis with data on prey abundance. Similarly, aquatic prey abundance can be analyzed as a trend; as input into a spatially explicit model (ATLSS); or as the dependent variable in a multivariate regression analysis with nesting success. Together, these analyses form lines-of-evidence (LOE) that contribute to our understanding of how well the hypothesis captures the functional relationships between the variables. However, there remains the issue of synthesizing and interpreting the information from each LOE to inform a decision.

6.6 Environmental Indices

Indices are widely used to convey clear and simple messages about what is happening to the environment to decision/policy makers and the public at large (www.wwf.org; www.heinzctr.org, www.epa.gov). Most indices are built from environmental indicators and serve to: (1) reduce the number of measurements and parameters that normally would be required to give an "exact" presentation of a situation; and, (2) simplify the communication process by which the results of measurement are provided to the user. Indices are included here as one more option for integrating the diverse suite of variables and multiple lines of evidence that form the basis of specific CERP hypotheses regarding the restoration of the South Florida ecosystem.

An index can be defined as an aggregated and weighted assemblage of one or more like elements that can be used to track the overall performance of these elements. For example, economists use the Dow Jones Industrial and S&P 500 indices to follow the performance of the stock market. Each of these indices represents a sub-set of stocks in the total stock market. If your interest is in an aggregate measure of large capitalization stocks then the S&P 500 index will provide that information. However both within the index and outside (stocks not included as part of index) there can be stocks that have totally different trajectories (i.e., rates of growth). If knowing the performance of one specific stock is important to you the index will not provide you that information. Consequently, both the composition (scale) and degree of aggregation used to develop the index must be relevant to the questions needing to be addressed.

Applying this concept to CERP indicates that an index designed to assess the health of a specific module may be quite different than that designed for the total South Florida ecosystem. The same argument applies to performance measures – if we index them we'll get an aggregate view but if a particular property is critically important (e.g., dry-season recession rate) and it either isn't included or properly weighted in the construction of the index then something very important will not be properly evaluated.

Indices are being widely and successfully used for environmental management, however, there utility is dependent upon the management question being asked and whether you can adequately characterize the health of the system using a sub-set of data. Multi-metric indices (i.e., diversity, similarity, and biotic) have become the standard in the United States for accurately assessing watershed health (www.epa.gov/bioindicators). The index of biotic integrity (IBI) for freshwater streams includes: species richness and composition metrics; indicator species metrics; trophic function metrics, reproductive function metrics, and abundance and condition metrics.

There are five activities that are central to constructing effective multi-metric biological indexes. The first is classifying environments to define homogeneous sets within or across similar habitat types (e.g., streams, lakes, or wetlands; large or small streams; warm-water or cold-water lakes; high- or low-gradient streams). The second is selecting measurable attributes (i.e., metrics, performance measures, etc.) that provide relevant and reliable signals about the ecological effects of stressors. The third is developing sampling protocols and designs to ensure that those physical, chemical, and biological attributes are measured accurately. The fourth is devising analytical procedures to extract and understand relevant patterns in those data. The fifth is communicating the results to the decision-makers, policy-makers, and the public.

6.6.1 Aggregation of Variables

The aggregation of variables is crucial to developing an index that accurately reflects the state of the ecosystem. The term aggregation is used to refer to the grouping and amalgamation of two or more different variables into one index. It is not used to refer to the grouping of the values of the same variable at different sites to achieve some overall representative value of that variable over a larger area. Nevertheless, most indices are also based on some kind of geographical aggregation that requires classification systems, definitions, nomenclatures, data production methodologies, measurement methods, etc., be consistent among the contributing data sources. One generally distinguishes:

- Spatial aggregation that is dependent on geographic scale. Regional and national environmental statistics are often based on the compilation and aggregation of data produced at sub-national level. The choice in geographic scale influences the area over which monitoring results can be estimated and whether the data can be aggregated on an ecosystem, administrative boundary or other geographic level and be representative of conditions over that area.

- Temporal aggregation is linked to the natural "variability" of the parameters monitored and to the need for more synthesized and usable information (e.g. annual averages for parameters measured daily or even hourly).

- Thematic aggregation is linked to the need for more readable and digestible information. Thematic aggregation establishes totals based on data for subcategories (e.g. total SOx emissions based on emission inventories, or total water resources based on water accounts). It may further be used to establish indices of urban air quality, global warming potential, acidifying substances, nutrient balances, etc., through the use of proper conversion factors

6.6.2 Aggregation Methods

Aggregation can be defined as the process of combining variables (e.g., physical, chemical, and biological performance measures) or units with similar properties to produce a single metric that represents the approximate overall value of its individual components. Even though each of indices may use its own unique aggregation methodology, the aggregation of two or more variables into one index typically involves the following steps: 1) selection of variables that are representative of the topic, policy issue or phenomenon of interest; 2) transformation of data is necessary when the selected variables do not have the same dimension ("apples and oranges") and to ensure that changes in one variable do not dominate those of the others in the final score of the index. This requires the use of a common metric as well as normalization and/or standardization to produce a common scale.); 3) weighting of the constituent variables before combining them, and 4) valuation of index scores by comparing them with a predetermined classification of what constitutes acceptable or unacceptable values.

The use of indices is offered here as one more approach to the integration of multiple-lines of evidence for the purposes of informing decision-makers and the public as to the status and trajectory of the South Florida ecosystem restoration.

6.7 Summary

The integration and interpretation of the impacts of multiple physical and chemical stressors on a specific ecological attribute, as described in a CERP hypothesis, will require the application of quantitative statistical and modeling methods as well as integration methods such as multiple-lines-of evidence (LOE) and weight-of-evidence (WOE) approaches. There is no "silver bullet" or single model that can address the complex issues of the temporal and spatial scales that characterize CERP. The intent of this guidance has been to present a very brief overview of statistical, modeling, and WOE approaches that have been used to analyze, integrate, and interpret monitoring data for the purposes of inferring the relationships between the stressors and effects described in the MAP hypotheses. Given the limitations of the various analysis approaches discussed above, it is evident that combinations of statistical, modeling and LOE/WOE methodologies will be needed to evaluate the MAP hypotheses and provide the information needed for interim goals and adaptive management.

6.8 Additional Applicable Resources

Human and Ecological Risk Assessment (December 2002, Volume 8, Number 7). Weight of Evidence Debate and Commentary. While papers in this volume focus on the application of WOE approaches and assessments of sediment quality issues the concepts and methodologies are applicable to or can be tailored to CERP assessments. It is a useful resource on the topic.

Human and Ecological Risk Assessment (December 2003, Volume 9, Number 1). Causal Relationships between Exposure (Stressors) and Effects in Field Studies: Debate and Commentary. There are 14 papers in this discussion that are relevant and can be tailored to making an assessment of the MAP hypotheses.

Details on the development, comparisons of approaches, applications at various scales, examples from different habitats, as well as recommendations of useful statistical methods for constructing and analyzing multi-metric indices can be found at the following websites. (www.epa.gov/bioindicators, www.wwf.org, and www.heinzctr.org/ecosystems/report.html)