A FRAMEWORK FOR ASSESSING THE EFFECTIVENESS OF PROGRAMS AND OTHER MEASURES DEVELOPED TO ADDRESS THE OBJECTIVES OF THE GREAT LAKES WATER QUALITY AGREEMENT

Submitted to the International Joint Commission by

James. P. Hill and Daniel Eichinger
Central Michigan University
March 8, 2013

Abstract

This report provides a framework for assessing the effectiveness of programs and other measures conducted under the auspices of the Great Lakes Water Quality Agreement (GLWQA), as amended in 2012. The report begins with a literature review of effectiveness evaluation in general, as well as an exposition of the specific challenges of assessing the effectiveness of environmental programs. It identifies and explores the requirements for developing an effectiveness framework for the Great Lakes, and then analyzes four potential models/approaches for the International Joint Commission to consider in performance of its assessment responsibilities under the GLWQA.

The authors in this report recommend a quantitative-qualitative effectiveness framework, based on an Elbe River case study which has many relevant similarities to the Great Lakes region. They then develop a proposed Great Lakes Environmental Effectiveness Metric (GLEEM)to assess the comparative effectiveness for programs and other measures in meeting both GLWQA Objectives as well as assessing the effectiveness of programs nested with each individual Objective.

The report concludes with an invitation to the Commission to apply aspects of the other three effectiveness models presented to future program effectiveness assessment efforts, providing brief summaries of these models as well as copies of full case studies in the report appendices.

Background and Overview of Report:

The Great Lakes Water Quality Agreement (GLWQA), as amended in 2012, represents an important, binational commitment between the governments of the United States and Canada to cooperatively work to "restore, protect, and enhance the water quality of the Great Lakes." Under Article 7 of the GLWQA, as revised in 2012, the International Joint Commission is charged with the responsibility of:

> analyzing and disseminating data and information about the General Objectives, Lake Ecosystem Objectives and Substance Objectives, and the operation and effectiveness of the programs and other measures established pursuant to this Agreement. Article 7 (1)(b)

The Commission is also charged in with:

> Providing to the Parties, in consultation with the Boards established under Article 8, a triennial "Assessment of Progress Report" that includes:
>
> > (iii) an assessment of the extent to which programs and other measures are achieving the General and Specific Objectives of this Agreement. Article 7 (1)(k)(iii)

On September 10, 2012, the International Joint Commission entered into a contract with Dr. James Hill (the contractor) of Central Michigan University to develop a framework for assessing the effectiveness of programs and other measures under the Great Lakes Water Quality Agreement.

The contractor was charged with the following tasks:

1. Conduct an extensive literature review and perform research to catalogue representative examples of various Canadian, U.S. and appropriate international

methodologies for assessing government program effectiveness. The methodologies presented will be those considered by the contractor to be appropriate and useful in order to help the Commission meet its obligations under the revised Great Lakes Water Quality Agreement, and shall include methodologies used in program assessments undertaken primarily by federal, state, and provincial units of government and internationally used methodologies

2. Provide a draft report (due January 15, 2013) which will include the following:
    a. Appendices that describe case studies and/or examples of methodologies applied in the program effectiveness assessments mentioned above as well as copies of summaries of the assessments themselves.
    b. Results of interviews and/or additional data, as needed, to strengthen the contractor's final recommendations and report as it relates to appropriate methodologies to address the Commission's responsibilities under the GLWQA. .
    c. Analysis and discussion of the strengths and weaknesses of each methodology/approach.
    d. Analysis of the advantages and disadvantages of applying each methodology/approach to the Commission's intended assessment.
    e. Recommendations (in rank order if desirable) of an assessment framework to the Commission, taking into account the Commission's mandate under the Agreement to supply "an assessment of the extent to which programs and other measures are achieving the General and Specific Objectives of this Agreement" including Agreement annexes as part of its triennial assessment of progress report.

3. After IJC staff and external review and comment (already completed), the contractor will submit a final report by March 18, 2013.

The contractor has provided a brief summary of what appear to be among the most promising methodologies to address the Commission's GLWQA responsibilities at the end of this report just prior to the reference pages, a representative sample of the most relevant national and international case studies in the appendices of this report, and most significantly an effectiveness framework from which can be derived an effectiveness metric to assess the diverse programs and other measures employed under the auspices of the Great Lakes Water Quality Agreement, as revised in 2012.

What follows is a summary of the research, finding, and conclusions reached by the contractor. An invaluable outside resource, employed according to the terms of this contract to provide important research and additional insight into this project, was Daniel Eichinger, a knowledgeable, well-qualified, and experienced Great Lakes researcher. His work on this project was significant and thus he is listed as a co-author of this report.

This report begins with an overview of the evolution of program effectiveness studies and research from the general evaluation literature and then proceeds to review the more recent literature and reports on environmental program effectiveness. It culminates in a working definition for environmental program effectiveness, which is used as the basis for many of the conclusions and recommendations drawn in this report.

The report next addresses the various typological methods of assessing environmental program effectiveness, concluding that the methodological approaches of Langbein and Felbinger's Causal-Output/Outcome Type and Chen's Outcome-Assessment evaluation constitute the most appropriate framework for this project.

Following discussion of the importance of having high quality and useful environmental data and indicators among other environmental assessment considerations, the report outlines four key features for any environmental model chosen to assess the effectiveness of key programs identified in the Great Lakes Water Quality Agreement. We conclude that the model considered to be the most valuable to the Commission in fulfilling its effectiveness responsibilities should be retrospective, causal, normative, and cross disciplinary.

Next, we describe several models that our research led us to conclude would be most useful in a Great Lakes environmental program effectiveness study conducted by the Commission. Discussions of the advantages and disadvantages are also provided, followed by our recommended framework -the No-Regime Counterfactual/Oslo-Potsdam Solution model and its application in the Elbe River case study. We shall entitle this application to the GLWQA as the **Quantitative-Qualitative Effectiveness Framework (hereinafter referred to as the QQEF approach.** The effectiveness metric we use in this framework we term a "**GLEEM**" score, which is an abbreviation for a Great Lakes Environmental Effectiveness Metric Score that can be used to measure comparative effectiveness.

While each of the models presented in this report have promise and potential challenges, the vast Great Lakes region and the breadth of programs that the IJC has been

charged to assess in terms of effectiveness is so diverse as to defy a one-model-fits-all-definition. Indeed, there is much to be gained by combining parts of several models presented in this report as part of an overall program effectiveness model, depending upon the context and complexity of the issue. For example, we discuss the use of a conservation audit described as model 4 of this report as it would be an appealing addition to comparing internal program effectiveness within a single GLWQA Objective.

However, the primary objective of this report is to develop a framework within which assessment of the effectiveness of GLWQA programs and other measures can be undertaken. The non-QQEF models perform contrasting frameworks to this preferred approach, serving both as a supplementary approach as well as offering future application values.

Introduction to Effectiveness Literature

We begin with a description of the current state of effectiveness evaluations for environmental programs. Our approach to this project focuses on four key tasks:

(1) Provide a review of the literature derived from evaluation research, government reports and case studies which establishes the role of effectiveness evaluations as a component of normative program evaluation.
(2) Provide a review of several major methodologies used to assess program effectiveness.
(3) Identify important limitations when conducting environmental program effectiveness.
(4) Present several different frameworks around which to organize an effectiveness review of U.S. and Canadian Great Lakes Water Quality Agreement-related programs and other measures, along with case studies that address issues of interest to the IJC and its assessment role under the GLWQA.

Defining Program Effectiveness

In order to develop a framework for determining the effectiveness of Great Lakes programs and other measures under the Great Lakes Water Quality Agreement, the contractor has been asked first to conduct a literature review on the development of government program effectiveness studies and reports. To do so we must begin with a discussion of the derivation of effectiveness studies from the program evaluation literature.

Formal evaluation of U.S. government programs evolved as a discipline during the 1960's as a component of Great Society social welfare programs. These evaluations focused upon

determining and documenting whether the policy objectives that these programs were designed to implement were indeed being implemented (Wholey, Hatry et al. 1994); (Langbein and Felbinger 2006). Despite its origins in the human services field, evaluation techniques have been widely applied throughout the social sciences. However, evaluating environmental programs is a comparatively new aspect of program evaluation, emerging in the middle to late 1990's .

The numerous definitions given for the evaluation process have resulted in this term being dubbed a 'semantic magnet' (Vedung 2000). Among them, Scriven (1991) provides this widely referenced definition: "Evaluation is the process of determining the merit, worth, and value of things."

Despite the widespread use of this definition, we prefer several other definitions that include judging effectiveness (Langbein and Felbinger 2006);(Owen and Rogers 1999), establishing causal links between a program and an outcome (Wholey et. al 1994), and retrospective assessment (Vedung 2000) because each, in turn, situates what we believe is a critical function of assessing program effectiveness in general, and are essential concepts in the service of this project. Based on our review of the literature and combining several of these definitions of evaluation, we propose the following working definition of effectiveness for this GLWQA project; namely,

> *the application of research methods (primarily in the social sciences and business) to retrospectively determine the causal links between a program and an outcome and judge the effectiveness of that relationship.*

Major Typological Approaches

In this section, we discuss, Scriven (1991), Langbein and Felbinger (2006) and Chen's (1996) approaches to identifying the major typologies in program evaluation that are useful in assessing program effectiveness. Scriven deconstructs evaluation into two types of evaluation: formative and summative.

Formative evaluation is designed to be used during program implementation to make adjustments in how the program is functioning, akin to what Blake refers [in Scriven (1991) and see also Chen (1996)] to as the "cook tasting the soup." The formative evaluation type is not designed to render judgment nor is it retrospective.

Summative evaluation is retrospective and is explicitly designed to render judgment on how successfully the program has achieved its intended outcomes. Extending Blake's metaphor, summative evaluation is akin to the "diners in the restaurant tasting the soup" (ibid). Scriven's typologies provide important coarse filter differentiation. However, to further narrow the methodological approach to program evaluation, we turn to Langbein and Felbinger (2006) and Chen (1996).

The methods and approaches to program evaluation are numerous and context dependent (Knapp and Kim 1998). Langbein and Felbinger (2006) identify four types of evaluation that are oriented around two distinct methodological approaches: descriptive and causal. Descriptive evaluation is more positivist in its approach with evaluation activities focusing on observation and measurement without judgment. Causal evaluation is more normative in its approach with evaluation activities focusing on detecting and considering relationships between a program

and a change in the goals the program is working to affect (Langbein and Felbinger 2006). Causal evaluation approaches result in a judgment.

Two substantive lenses condition Langbein and Felbinger's approach to program evaluation: implementation and output/outcome (2006). The implementation lens looks to answer mechanistic questions about how the program functions where the output/outcome lens assesses the results of the program on its intended outcomes, which is consistent with the Commission's task under GLWQA.. Langbein and Felbinger pair methodological approaches in Table 1 with substantive lenses to create four typologies for evaluation, each answering related, yet different questions:

Table 1: Four Evaluation Typologies

Typology	Key Question
Descriptive-Implementation Evaluation	How was the program organized?
Descriptive-Output/Outcome Evaluation	Have program efforts focused on the right questions?
Causal-Implementation	What caused the program to select this approach to implementation?
Causal-Output/Outcome (**Recommended)**	Did the program achieve its intended outcomes?

Source: adapted from Langbein and Felbinger (2006)

Chen (1996) also identifies four types of evaluation (see Table 2)

Table 2: Four Types of Evaluation

	Evaluation	
	Improvement	Assessment
Process	Process Improvement Evaluation	Process Assessment Evaluation
Outcome	Outcome Improvement Evaluation	Outcome Assessment Evaluation (**Recommended**)

Source: Chen (1996)

As in Lanbein and Felbinger (2006), each type elaborates upon a different aspect of the evaluation function. Process improvement evaluation is used in the style of Scriven's formative assessment not to provide overall judgment of the program, but to identify opportunities to strengthen elements of the program (Chen 1996). Chen likens process assessment evaluation to "quality control" which judges whether the implementation of the program was a success or failure. The process perspective, like Langbein and Felbinger's implementation approach is concerned with *how* a program functions.

Outcome improvement evaluation stops short of evaluating overall effectiveness but is used to differentiate which activities within a program are more or less important for successful implementation (Chen 1996). Outcome assessment evaluation, unlike the other types, is designed to provide an overall judgment on whether and how well the program has met its stated goals. (Chen 1996). This approach is the typological cousin to Scriven's summative evaluation type and Langbein and Felbinger's Causal-Output/Outcome type.

Chen also proposes that mixing types of evaluations can enhance the rigor of the evaluation approach. When the program is particularly complex, the evaluation must respond to questions that require different methodological approaches, or when there are diverse stakeholders for whom the evaluation is being conducted (Chen 1996).

Discussion of Methodological Approach

Thus, there are clear types of methodological approaches that result in answering different types of evaluation questions, depending upon the goal of the evaluation. Since the purpose of this project to assess the effectiveness of Great Lakes programs and other measures under the GLWQA, the methodological approach that is used to conduct the evaluation must necessarily render judgment on outcome achievement.

Thus, for this project, we believe that a methodological approach in the fashion of Langbein and Felbinger's Causal-Output/Outcome Type and Chen's Outcome-Assessment evaluation is the correct construct or framework, as it provides both causal-output oriented and an outcome assessment evaluation. It further provides a broad framework within which comparative and diverse effectiveness assessments can be undertaken.

Environmental Evaluations-Key Considerations

Having considered the major typological approaches to program evaluation, we now turn to key considerations for evaluating the effectiveness of environmental programs. However, we must begin with this cautionary statement from Knapp and Kim:

> "Environmental policy issues are characterized by scientific uncertainty; benefits that are diffuse and difficult to measure; costs that are concentrated, often large and easy to measure; and core values and beliefs that vary widely among the population. While there is hope that program evaluations can be both rigorous and well received, it is not clear that environmental policy is the natural domain for this to occur," (1998).

Despite this warning, conducting environmental evaluations and determining the effectiveness of a program in meeting its stated objectives is becoming an increasingly important part of environmental policy, commanding greater attention within the environmental community ((Mickwitz 2003);(Parrish, Braun et al. 2003); and (Hockings, Stolton et al. 2006).

As Knapp and Kim (1998) allude to in their quote, it is difficult to conduct environmental program evaluations that detect and discuss causal relationships between the presence of the program and a change in the measured outcome. Environmental problems are complex and the complexity is increased through the dynamic interaction of human, social, technical, and economic activities (Mickwitz 2003;Johansson 2006). In addition to these concerns, environmental problems can, at times, be observed only through a time-lag which confounds detection, attribution, and accurate measurement (Helm and Sprinz 2000).

The challenges of conducting program effectiveness assessments in other disciplines are greatly amplified when assessing the effectiveness of environmental programs; particularly the challenges of selecting the correct methodological approach; the availability of empirical data

(or lack thereof); and selecting the most appropriate indicators, (Mickwitz 2003; Knapp and Kim 1998; Helm and Sprinz 2000).

It is our view, based upon the preceding discussion, that conducting a meaningful and effective environmental program assessment depends to a great extent upon the availability of adequate and accurate data sources from which the causal relationship between the program and a change in the outcome can be detected and considered. The importance of this data cannot be overemphasized and will be an important factor in accurately measuring program effectiveness no matter what framework is selected. In short, one cannot judge program effectiveness without being able to first assessing the program effect on the outcome.

Environmental data or numerical indicators without outcome/causality linkage, however, is equally misleading. For example, on October 9, 2012, the U.S. Forest Service issued a press release declaring the Au Sable River Large Wood Restoration Project to be a "Success". This project was the culmination of a 10-year plan to reforest a ten mile stretch of the Au Sable River in the Huron National Forest (Michigan) by the planting the last 126 of 1200 trees transported by helicopter. The project was intended to restore the riparian system previously harmed by dams and logging.

The fact that the numerical tree replanting goal was completed begs the question of whether this numerical replanting goal resulted in restoration of the system. The success declared was in the process of replanting the trees by helicopter (an output), but achieving this goal does not necessarily translate into the primary objective of restoring the river system below the Alcona Dam - an outcome assessment much more demanding in terms of data and causality.

Significant sources of Great Lakes environmental data exist from the EPA's GLAS data base (such as their measures of progress for the Great Lakes Restoration Initiative) and also from State of the Lake Ecosystem Conference (SOLEC) indicators. The GLAS system describes what Great Lakes Restoration Initiative (GRLI) projects have been funded, who received the money and how it is being spent, and most importantly the progress being made toward achieving the goals of the GLRI. Simple and useful measures of progress are included in the GLRI users guide for toxic substances, AOCs, invasive species, nearshore health and nonpoint source pollution, and habitat protection and restoration. The SOLEC indicators assess the state of the Great Lakes ecosystem based on accepted indicators and provide an historical, quantitative basis for effectiveness assessment purposes.

However, lack of quality Great Lakes data due to a lack of government research funding and insufficient monitoring, as well the need for better linkage of SOLEC indicators to appropriate GLWQ Objectives raised in government reports such as the IJC's 13th Biennial report (December 2006) are of concern to the authors of this report. Whatever framework the Commission chooses to pursue to meet its Great Lakes assessment responsibilities under the GLWQA, the success of the chosen approach depends upon sound environmental data and core indicators. As a preferred framework is adapted and applied, there must be a high priority placed on ensuring that the data and core indicators used in the framework are reliable, longitudinal and linked to the Objectives of the GLWQA.

Evaluation Models

In this section, we focus on providing more descriptive discussions of several different models for conducting an environmental effectiveness evaluation appropriate for assessing GLWQA programs and other measures. In doing so, we return to the definition of effectiveness evaluation we constructed previously; namely:

> *the application of research methods (primarily in the social sciences and business) to retrospectively determine the causal links between a program and an outcome and judge the effectiveness of that relationship.*

We also now apply the typology we discussed earlier as well. By defining evaluation and discussing the major typologies, we have winnowed our effectiveness focus in order to identify and present evaluation models that:

(1) are retrospective,

(2) result in the determination of causal linkages, and

(3) make judgments about the program relative to its stated outcomes.

A fourth, and in our view essential feature of these models as they relate to the Great Lakes is that they must also generally:

(4) be applicable across a variety of disciplines.

A discipline-specific evaluation technique has value when comparing like programs. However, In this project, we are seeking to evaluate program effectiveness across a wide range of Great Lakes programs (both in scope and geography) and disciplinary perspectives. For example,

program effectiveness can borrow techniques from the business perspective as well as scientific techniques from a biological perspective.

The framework most useful for this project should be one that renders these different programs comparable to one another with an objective measurement of effectiveness. For this reason, many published environmental program effectiveness studies are not discussed in this report, as they were constructed as essentially single serve evaluations. However, in the WCPA framework discussed later in this report, we note that these single serve evaluations often include steps that can be useful for measuring effectiveness.

Accordingly, we identify for consideration the following four models that best meet our previously mentioned criteria of being applicable to many of the GLWQA issues and have been used in international contexts as well as by government agencies and environmental NGO's.

Model 1: The No-Regime Counterfactual/Oslo-Potsdam Solution (Preferred Framework)

The No-Regime Counterfactual/Oslo-Potsdam Solution (NRC) model is described by Helm and Sprinz (2000) and Hovi, Sprinz et al. (2003). The NRC is a combination of two methods proposed by Underdal (1992) for assessing regime effectiveness, and has been adapted by Helm and Sprinz (2002) and Hovi et. al (2003) for determining the effectiveness of environmental regimes. This model has been used by Helm and Sprinz (2002) to evaluate regime effectiveness in Europe for responding to transboundary air pollution problems and by Dombrowsky (2008) to evaluate the effectiveness of environmental conservation programs on the Elbe River.

The NRC works by selecting a lower bound which is defined as the no-regime counterfactual, a figure which posits what the measurement for a given indicator would be in the absence of the regime/program. The upper bound, which is defined as the collective optimum, posits what the measurement for a given indicator would be under ideal and unconstrained conditions. There is often a lack of truly comparative data from which to derive the figure for both the no-regime counterfactual and the collective optimum. To address this lack of comparative data issue, Helm and Sprinz (2000) and Miles, Underdal et al. (2002) use standardized, structured, expert-based scoring mechanisms for determining point estimates for the no-regime counterfactual.

For selecting the collective optimum, Miles et. al uses a structured, empirical approach, derived from external data sources to determine the collective optimum(Hovi, Sprinz et al. 2003) .

We believe for this Great Lakes program effectiveness report that the same approach used to determine the no-regime counterfactual can also be used to determine a point for the collective optimum, a point we will develop later in this report on our recommended framework. In this regard we depart slightly from the literature: Helm and Sprinz (2002) use Nash's equilibrium game theoretic model to determine boundary points for the collective optimum within a game-theory context that could predict environmental treaty adherence among nations. For this project, when assessing program effectiveness, the game-theory context is not applicable and using Nash's equilibrium, as Helm and Sprinz do, offers no real benefit to the construct we propose. Similarly, the regression difference in difference (DID) design devised by Vollenweider (2012) when studying long range transboundary air pollution would not be appropriate for devising a convincing counterfactual for this model.

Below in Table 3 is a diagram of how the NRC model would work.

Table 3: The NRC Equation

There are several different ways to express the NRC:

$$E = \frac{AP - NR}{CO - NR}$$

E: Effectiveness

AP: Actual Performance

NR: No-regime counterfactual

CO: Collective Optimum

This basic equation shows the relationship between the upper and lower bounded values. In the numerator, we have the actual measured performance of the program on a given indicator and calculate the distance between the actual performance and the no-regime counterfactual. This demonstrates the level of effectiveness of the program against what would have occurred in the absence of the program. In the denominator we measure the distance between the collective optimum and the no-regime counterfactual. Expressing the model in this fashion calculates the numeric effectiveness (0 to 1 scale) relative to the amount of effect that could have been observed. This approach provides several important advantages:

(1) It measures the observed effect and relates that measure to the amount optimal effect that could have occurred.

(2) The model is not biased towards either the upper or lower bound as it could be if only one of the relationships was calculated.

(3) Because the result of the evaluation is a readily accessible value, it can be easily understood and the effectiveness of diverse programs can be measured and compared.

There are, of course, other considerations in the use of the NRC. It is important to be aware of Young's (2001) critique of counterfactual analysis that determining the counterfactual values, using qualitative and normative techniques, is inherently problematic depending upon the methodology employed (Young 2001). Hovi et. al acknowledge Young's critique but do not consider it to be fatal (Hovi, Sprinz et al. 2003).

This model has been applied to review the effectiveness of the International Commission for the Protection of the Elbe River (ICPER) (Dombrowsky 2008) See Appendix A and a brief summary of the study at the end of this report. The ICPER maintains areas of program activity that are similar to those of the IJC in that they focus on municipal wastewater, industrial point source pollution, agricultural non-point source pollution, contaminated sites and landfills, fish migration, protected areas and morphology, and accidental pollution.

As Dombrowsky demonstrates, this model blends qualitative and quantitative research methods to provide an evaluation of overall effectiveness as well as effectiveness at the program level. Dombrowksy also makes use of the structured, systematic scoring technique to detect and determine the causal link between program activity and observations on the outcome.

Dombrowsky concludes in the Elbe River study that:

> the quantitative approach of the Oslo-Potsdam solution for measuring effectiveness provided analytical clarity and contributed towards showing the different levels of effectiveness in the different areas of activity. At the same time, the qualitative approach (structured expert interviews) contributed towards a better understanding of causal relationships.

Young (2011) dubs this approach as "ambitious" in terms of its combined quantitative and qualitative analysis. However, the completeness of this model, in terms of its fulfillment of our previously established criteria for an effectiveness measure, makes it our preferred framework, as we will outline in the implementation section of this report.

Moreover, this framework could be powerfully combined with models from other typologies to also address diagnostic questions about what characteristics or procedures make a program more or less effective. We believe this model could be a useful one when applied to many of the priority programs under the GLWQA. (See the **Elbe River case study in Appendix A)**

Model 2: WCPA Framework

The second model to consider is one developed by the International Union for the Conservation of Nature's (IUCN) World Commission on Protected Area (WCPA). In 1997, the WCPA convened a Management Effectiveness Task Force to design a framework to assess management effectiveness for parks and protected areas.

The framework, which has since been adopted as a best practices guide, proposes three fundamental levels for conducting evaluations (Hockings, Stolton et al. 2006) **See Appendix B**

for an exposition of the park managers application. The three levels ascend in complexity and appetite for data:

Level 1 is an inferential evaluation process that uses existing sources of data to make determinations;

Level 2 combines the approach in level 1 with limited monitoring of outputs and outcomes; and

Level 3 emphasizes monitoring the extent of achievement of management objectives by focusing on program output and outcomes, while retaining measures of context, planning, inputs and processes (Hockings et. al 2006).

Level 3 evaluation is the most intensive evaluation level under this model, demanding the most data. Table 4 below elaborates upon elements of this evaluation.

Table 4: Elements of Evaluation

Elements of Evaluation	**Explanation**	**Criteria that are assessed**	**Focus of evaluation**
1. Context	**Where are we now?** Assessment of importance, threats, and policy environment	significance, threats, vulnerability, national context	Status
2. Planning	**Where do we want to be?** Assessment of protected area design and planning	protected area legislation and policy, protected area system design, reserve design, management planning	appropriateness

3. Inputs	**What do we need?** Assessment of resources needed to carry out management	resourcing of agency, resourcing of site, partners	resources
4. Processes	**How do we go about it?** Assessment of the way in which management is conducted	Suitability of management processes?	efficiency and appropriateness
5. Outputs	**What were the results?** Assessment of the implementation of management programs and actions/delivery of products and services	results of management actions, services and products	effectiveness
6. Outcomes	**What did we achieve?** Assessment of the outcomes and the extent to which they achieved objectives	impacts: effects of management in relation to objectives	effectiveness and appropriateness

In this model, conducting an outcome evaluation depends upon determining how goal attainment and overarching values were defined at program commencement. Monitoring the status of those values includes selecting key attributes of the value; identifying appropriate indicators, and selecting a methodology for assessing the indicator- important features in any framework selected by the Commission.

Parks Canada used this method to assess program performance on the following values: Is the park losing native species? Are selected indicators within acceptable range? Are herbivores and predators playing their role? Are biological communities at a mix of ages and spacing that will support native biodiversity? (Hockings, Stolton et al. 2006). This model represents a

standardized approach to the effectiveness of Parks Canada protected area management that can be applied to solve for the challenges of evaluating vastly different biological communities.

The advantage of this model is that it represents a fairly standard approach in program evaluation in general; namely, understand what was intended; determine the values that carry that out; select a proxy for that value to measure, and then conduct the measurement.

We also count among the advantages of this model its ability to standardize evaluation, but only within effectiveness evaluations of similar type. In other words, evaluating protected area to protected area or coastal management program to coastal management program is both its strength and its limitation. Therefore, the application of this model is useful in evaluating within a major program category like habitat and wildlife protection, an important and a difficult activity to assess in terms of program effectiveness when compared to measuring more easily measured pollution reduction programs. Nevertheless, it may very well serve as a benchmark for evaluating particularly difficult-to-assess GLWQA priority issues that are more subjective in nature, and that alone may merit its consideration by the Commission as part of its effectiveness arsenal.

However, this model does not adapt well across other difficult categories like nonpoint source pollution. We also are concerned that the model does not attempt to explicitly probe the causal link between recorded observations of the indicator and the program.

As we have discussed previously, one of the key challenges in environmental program evaluation is how to be certain that the observed changes are due to the program being

evaluated or some other confounding variables. It has the same problem as using core indicators on a pre-post test basis- it is hard to find a direct causal basis. As Sprinz (2000) summarizes earlier researchers concerned about the problem of using such stand-alone tests or GLAS indicators:

Simple pre-post tests (i.e. focus on the change in a performance variable before and after an intervention) clearly fail to establish a compelling quasi-experimental design- because they cannot rule out that an unobserved variable caused the change in the performance variable.

We are also concerned about the thin description for methodically approaching the selection of attributes and indicators. A standard approach or methodology for systematically selecting these indicators under this model is important in forming a valid basis for comparing programs.

Lastly, this approach relies exclusively on a qualitative approach to evaluation. While we make no argument in preference to quantitative over qualitative methods, we believe that using techniques from **both** methodologies (like the NRC model) provides an important measure of confidence. Issues of subjectivity and frames of reference can weaken qualitative measures.

Model 3: Conservation Excellence Model

A third model to consider is one developed by Black and Groombridge (2010), who propose an adaptation of the European Foundation for Quality Management's (EFQM) business excellence model for application in a conservation/environmental context. The primary focus of this model is to assess organizational effectiveness at achieving its outcomes (Black and Groombridge 2010). The conservation excellence model deconstructs the evaluation along 9 approach criteria, similar to the 9 box model used in the EFQM business excellence model.

Table 5: Conservation Excellence Model: Approach and Results Criteria

Approach	Sub-criteria
1. leadership	a. leaders demonstrate commitment to conservation b. provision of resources or assistance (finance, people) c. direct involvement with conservation organizations and stakeholders d. recognition and encouragement of efforts, achievement, and ideas
2. Policy and strategy	a. policy, strategy, and plans use relevant, comprehensive information b. policy, strategy, and plan development involves relevant people c. Policy and strategy are effectively communicated and implemented
3. people and Local community management	a. planning and improvement of people resources (workers, volunteers) b. people's capacity is sustained and developed (training, education) c. people agree on targets and review results (in project) d. people are involved and empowered (roles, decision making, rights) e. people have effective communication and decision making) f. well-being of people planned, managed, and monitored
4. resource management	a. financial management (budget, accounts, records, authorization) b. information management: access, structure, validity, security c. supplier of materials management (selection, contracts, storage) d. buildings, equipment, and asset management (maintain and use) e. intellectual property (relevant information used and protected)
5. core conservation Processes	a. core processes identified systematically on research basis b. processes and responsibilities managed systematically c. processes reviewed (technical results, adaptive management) d. processes improved through innovation, creativity e. processes improved (change implemented, monitored, evaluated)

Results:	
6. biodiversity Results	a. biodiversity response to program actions (habitat, population, range) b. other measures (e.g. ecosystem function, geophysical measures)
7. people and Local community Results	a. staff and community perceptions of the program (e.g. via surveys) b. other measures (community involvement, conflict, well-being)
8. impact on Wider audience	a. perception of wider society (awareness, attitudes, political support) b. indirect measure (threats, legislation, donations, volunteers, press)
9. conservation program results	a. financial measures of success (income, funds, investment, budgets) b. non-financial measures (program targets achieved, milestones)

The conservation excellence model can be applied in a variety of ways depending upon the purpose of the evaluation, including rating a program(s) to evaluate overall effectiveness or making comparisons across programs, including post-program evaluation (Black and Groombridge 2010). The evaluation steps are listed in Table 6.

Table 6: Evaluation Steps

Evaluation Steps:	
1. Collate and assess data	Collect and assess data against the four results criteria.
2. Assess policy and strategy criteria	Assess whether policies and strategies address the program issues raised in the step 1

3. Identify and review core conservation processes	In terms of quality and innovation of approach
4. Review management of workforce, communities, and resources	In terms of capacity and involvement of people, allocation of funds, and use of assets achieve program objectives
5. Assess leadership	In relation to activities occurring in the organization and whether correct actions are (or not) being reinforced by leaders

The conservation excellence model relies upon a five-step assessment process to collect the data sources upon which to formulate different evaluations. In using this model for effectiveness evaluation, Black and Groombridge (2010) review the results first and trace the evaluation criteria and sub-criteria back from that data point. (**See Appendix C for this model in the context of species conservation programs)**

In our view, this model has several advantages.

(1) The criteria and sub-criteria are comprehensive and correctly link institutional performance with outcome achievement.
(2) The five-step approach is adaptable using a variety of data collection techniques.
(3) The model can be used to conduct retrospective effectiveness evaluation (as Black and Groombridge demonstrate).

The model falls short, however in that it assumes a causal relationship between the activities of the organization and measured or observable changes in outcome, a common problem

identified in causal inference literature, which is concerned with the impact of unobserved intervening variables. Furthermore, the model strays across typologies by focusing some of the steps on making diagnostic calls about program implementation issues.

If this model is selected by the IJC, we would propose coupling its techniques and approaches with another model that more explicitly seeks to establish a causal link between the program and any change in the outcomes, perhaps with the addition of more qualitative inputs from inside and outside experts. Furthermore, while this model features what could be a stand-alone evaluation technique, it was clearly designed, as was the next model we describe, for planning conservation programs.

As this model has only recently been published (2010), we are unaware of published accounts of its application. However adapting business evaluation models to environmental issues is not a new phenomenon. For example, Bronson and Noble review the effectiveness of Park Canada's use of ISO 14001 environmental management systems. (Bronson and Noble 2006) . In any event, this model does have possibilities for assessing and improving GLWQA program effectiveness.

Model 4: Conservation Measures Partnership-Conservation Audit

We offer the conservation audit as a possible fourth model, although we see it more as a supplementary tool for broader effectiveness frameworks that could strengthen programs within an Objective. The Conservation Measures Partnership (CMP) is a consortium of conservation NGO's that have assembled what they term are the best practices for carrying out

conservation programs (CMP 2007). The CMP essentially proposes an adaptive management framework that features what they view to be the fundamentals of successful planning.

In and of itself, this approach is not particularly useful for this project. However, we believe there is much value in its approach to supplement planning conservation programs. What we found compelling was their use of what they termed **conservation audits** to review the planning, execution/implementation, and results of a conservation project or program (CMP 2007). Member organizations of the CMP had (as of their 2007 report) conducted 37 conservation audits at varying scales including for global programs and multi-national level conservation programs.

The audits been applied as an extension of the adaptive management loop to determine how faithfully conservation projects adhere to the best practices designed by the CMP. The audit is conducted by a project team with a designated team leader who assembles the needed expertise [**See Appendix D for example of a conservation audit of Shiawassee River Watershed Project (Michigan)].**

Different survey and assessment instruments are used to probe the different actions taken by program leadership, including a self-assessment and external assessment of their performance against the best practices. The audit results have been used in the formative rather than summative sense, which is to say, that they have informed adjustments or changes to how the program is being delivered, but not to render a summary judgment on the overall effectiveness of the program-which is the purpose of this project.

While the typology and approach of this technique do not conform to what we believe are called for in this project that is not to say that the "auditing" framework would not contribute to an effectiveness review. One of the distinct advantages of this approach is its participatory nature. This method would be particularly effective as a technique to help determine causality- not unlike the methods discussed in the NRC model.

We do not favor this model as a stand-alone framework. Rather, we present it to highlight some of the novel approaches used in this method and the value it adds to other models described previously.

Machmer and Steeger (2002) developed an effectiveness evaluation protocol originally used in British Columbia's Terrestrial Ecosystem Restoration Program (now defunct) which includes familiar approaches to the conservation audits. The evaluation method they describe is flexible enough to apply to a variety of ecosystem applications but is not explicitly retrospective as our definition requires. However some adaptation of this approach to include retrospective evaluation is possible and of use in measuring the difficult issue of determining the long-term, comparative effectiveness of an ecosystem restoration program.

A copy of the final report submitted to the British Columbia government is included as **Appendix E,** though it is not formally evaluated for purposes of this report.

Recommendations:

Framework Explanation/Implementation Plan for the Great Lakes

We have now outlined several potential frameworks which, when applied to a specific policy, program, organization, or regime, may result in establishing some measure of effectiveness for the Great Lakes. However, our preferred methodological approach is the Oslo-Potsdam solution (NRC) as outlined in the Elbe River case study previously discussed and provided in Appendix A. **We have renamed this approach the Quantitative- Qualitative Effectiveness Framework or QQEF.**

One of the key advantages this QQEF approach is its use of both quantitative and qualitative measures, adding richness and diverse internal and external expert input into the equation. Another distinct advantage to this approach is that it provides a broad framework by which dissimilar environmental problems can be compared to one another. The breadth of topics under the Great Lakes Water Quality Agreement demands an effectiveness measure that places the complementary but distinctive Objectives in relation to one another in comparative metrics.

The process we outline below tracks closely the steps pursued in Dombrowksy's work on the Elbe River. While we follow this model closely, there also are opportunities to select additional/alternative approaches to the steps Dombrowsky follows, which are noted in the discussion that follows. Thus, we provide an explanation of the essential steps involved with implementing this framework for the Great Lakes, including:

(1) choosing indicators

(2) assembling the data collection tool

(3) identifying expert interview/survey groups; and

(4) Conducting appropriate data analysis.

In addition, we have also constructed a scenario using hypothetical response data to demonstrate how to arrive at an environmental effectiveness score for Great Lakes programs. For abbreviation purposes, we will name this Great Lakes Environmental Effectiveness Metric a **"GLEEM**" score.

1. **Choosing indicators: the fundamental data task**

At the outset, it is essential to gather two pieces of fundamental data: the goals or objectives that are the purpose of the policy/program and the indicators that demonstrate the condition of those objectives. For the first task, we refer to Article 3 of the 2012 Great Lakes Water Quality Agreement (GLWQA) which outlines 9 General Objectives (Objectives) on which the agreement and the organizations working under the auspices of it are designed to work. The nine Objectives are provided in **Appendix F.**

It is important to note that any policy or program nested directly under each of these Objectives may also be evaluated for effectiveness. Thus, the result of a comprehensive effectiveness evaluation will be a report on the level of effectiveness between the Objectives as well as the programs within them. The process we lay out is essentially the same regardless of the level at which we are working.

Having identified the Objectives, we must also identify the indicator or sets of indicators which demonstrate the condition of that Objective. Given the numerous Great Lakes data vendors, there are a variety of sources from which to obtain indicator data. For this exercise, we shall refer to the indicator reports and summaries completed under the State of the Great Lakes report, which is posted on the EPA's website.

An important initial effort will be assigning different indicators to the various GLWQA Objectives. We make such an assignment in the hypothetical situation we have created later in this report. However, the actual assignment of indicators for the Great Lakes framework is a crucial responsibility better left to the expertise of the client and potential interview participant/survey respondents.

We also look to others to direct which data sources are viewed as having primacy for indicating the condition of a particular Objective. One such source might be the measures of progress described in the Great Lakes Recovery Initiative as part of the Great Lakes Accountability System (GLAS).

Data Collection Tool

There are two approaches that we propose could be used for collecting the data that will map onto our GLEEM calculation.

(1) First, as Dombrowsky demonstrates, is a **semi-structured interview technique**. Semi-structured interviews make use of a standard set of questions that are posed to the interview participant, yet provide flexibility to allow for follow up questions or for

pursuing threads of discussion that may be related to the topic, yet are not explicitly solicited via prepared interview question. Each interview would be recorded and a full transcription would be completed. The interview provides an opportunity for follow up and for seeking greater clarity between the interview participant and the interviewer, which can result in the collection of high quality data. We count among the disadvantages of this approach, the time consuming nature of conducting such interviews and the transactional costs associated with completing accurate transcriptions of the interview. Dombrowsky refers in his work to the relatively few number of interviews he conducted in his Elbe River study. A small *n* value can exacerbate a bias problem in the responses.

(2) A second data collection method, which differs from Dombrowsky, is to **solicit responses via survey**. The survey would pose the same questions as those presented in an interview context and would provide time for the interview participant to carefully construct and contemplate their responses. Among the advantages to this approach are the comparatively low transaction costs and the ability to reach out to greater numbers of participants because the constraint of time on the interviewer is lower. The major disadvantage is that the quality of the data *may* suffer without the ability to engage in the "unscripted" back and forth that could occur in the semi-structured individual interviews.

We favor the second approach for several reasons. We believe that reaching a broad number of respondents provides quantitative weight to the responses and will more accurately portray

variance among scores than a smaller number of more detailed responses. Our preference for this method should not leave the impression that the interview approach is any less significant in our opinion, simply more costly in terms of time and resources.

Prior to conducting the interview or beginning the survey, we would present each participant with an indicator summary (or perhaps GLRI measure of progress data from GLAS if more appropriate or relevant) for the Objective they have been asked to evaluate. This step introduces a common thread to the otherwise diverse backgrounds and experiences of the expert pool and would not seek to replace their parochial knowledge on the subject, but to offer a context around which to consider their responses.

The survey questions themselves are few and straightforward. To run the effectiveness evaluation, we must gather two pieces of information from the respondents: (1) a quantitative measure of the overall condition of the Objective; and (2) a qualitative assessment of the role GLWQA programs or other measures have played in contributing to that condition. We also recommend the inclusion of a third question on indicator appropriateness which serves as a confidence check on the indicators that were selected to represent that Objective.

We propose the following text as an example of what could appear on the survey instrument:

The following text summarizes the state of several indicators which have been selected to demonstrate performance with the respect to Objective X of the GLWQA [insert full text of Objective from Article 3]. After you have completed reviewing this summary(ies) please provide your responses to the following:

1. **On a scale of 0 to 10, how would you rate the level of accomplishment with respect to Objective (X) of the Great Lakes Water Quality Agreement?**

2. **Please identify and explain the specific contributions the programs or other measures of the GLWQA have made to this Objective.**

3. **In your opinion, do the indicators provided for this Objective accurately demonstrate the condition of that objective today? Are there other indicators that should be considered to better represent this Objective. If so is it available, and where may it be found?**

In an interview context, the questions will be asked in sequence without moving forward until the answer to the last question is completed. Similarly, if the questions are posed in an electronic survey, we would restrict moving forward until the previous question was completed, as prior knowledge of the second question may bias an answer for the first.

The third question is important as a confidence check that the underlying assumptions are correct about what we are observing with respect to a particular Objective. For example, if a SOLEC indicator is deficient in terms of relevance or quality, a GLAS derived measure of progress from the GLRI might be a better measure.

Expert Interviews

The choice of data collection methodology (interview versus survey) will determine the pool of prospective interview participants. We believe that the Commission should be deeply involved in identifying the organizations and individuals with expertise in assessing achievement of each of the Objectives identified in Article 3 from which responses can be solicited. In the Elbe River case study previously discussed, experts with both "insider" knowledge and "outsider" perspectives were sought, meaning that agency personnel were solicited in addition to personnel from related NGO's and stakeholders.

For this project, we propose identifying expert interview teams for each of the 9 objectives. For example, the expert pool for Objective (iii): "allow for human consumption of fish and wildlife unrestricted by concerns due to harmful pollutants" could include (but not limited to):

- Program staff from the EPA, Environment Canada, state and provincial governments who monitor pollutants present in sport-fish
- Community health managers responsible for evaluating consumption risks for humans and issuing consumption advisories
- US, Canada, state and provincial wildlife veterinarians who monitor wildlife health and exposure to pollutants
- Key public experts in the private and academic sectors

Data Analysis

The first interview/survey question directs the respondent to provide an overall assessment of the condition of that objective scoring between 0 and 10, which represents the value we assign for actual performance (AP). The responses will be averaged among all respondents to render a single score and to provide a check against bias.

We believe, as Dombrowksy does, that it is important to report on the variation among the responses by establishing a coefficient of variation for the responses. Coefficients closer to zero represent high consistency in the responses, and a coefficient closer to one represents greater variation in the responses.

The second question is more qualitative in nature in that it asks respondents to provide specific examples attributing the actual performance they scored in Question 1 with GLWQA programs or other measures. We do not condition how the response is given, nor does Dombrowksy, which means that all of the responses must be analyzed and categorized.

Dombrowsky undertakes this step on his own, analyzing the answers given by his respondents and assigning a qualitative assessment to them, which corresponds with a quantitative weight. The qualitative assessments and quantitative responses he uses are zero (0.00), low (0.15), low to medium (0.35), medium (0.50), medium to high (0.65), high (0.85), complete (1.0), with the quantitative weights for these assessments noted parenthetically.

As we note, Dombrowsky makes these assignments based upon his personal assessment of the responses. Without criticizing this approach, we propose a slight departure that will introduce some additional rigor to the process. We recommend assembling a small, independent team to code the qualitative responses. Each team member would charged with reading all of the responses given to question 2 and assigning a qualitative assessment to each response. The team would reconvene with the purpose of reaching agreement on the assessment for each response. We would then organize those responses and assign the quantitative weight for each response, average them, and calculate the coefficient of variation.

The quantitative weight is the adjustment calculation used to determine the no-regime counterfactual when it is applied to the actual performance (AP) figure determined under interview./survey question 1. In calculating the AP value, we take into consideration the

contributions of the GLWQA identified (the purpose of the second question) and identify what specifically is attributable to the programs and other measures of the GLWQA.

We then assign a significance to those actions, and adjust that value out of AP to render the NR, which is the state of the Objective absent the intervention of the policy or program.

The collective optimum (CO), or goal attainment, is always assigned a value of 10 in this exercise.

From these steps, we have assembled the data necessary for completing the effectiveness or GLEEM score by inserting the values into the equation identified in the Oslo-Potsdam Solution:

Where **GLEEM =(AP-NR)/(CO-NR) and where NR = AP-(AP*QW)**

Hypothetical Scenario

To provide a concrete application of this framework, we have constructed the following hypothetical scenario using Objective iii (fish and wildlife consumption) to demonstrate how the data would flow through this framework and results in the GLEEM score.

1. An expert pool is identified and surveys are sent out to 23 interview participants, 14 are returned.
2. For the first question the following responses are generated from the 14 respondents: 5.6; 3.7; 6.2; 6.7; 7.1; 4.4; 5.3; 4.4; 5.5; 6.2; 6.0; 4.8; 6.6; 5.9. From these responses we determine that the average score among the respondents is 5.6. This value is assigned as the AP. The coefficient of variation is 0.176.

3. In response to the second question, the 14 responses are independently coded by the small team. The team convenes and agrees upon the following qualitative assessments: 6-responses were coded as low; 6 responses were coded as low-medium; 2 responses were coded as medium. The quantitative weights were averaged against all responses. The mean quantitative weight is 0.2714; the coefficient of variation is 0.4999.

4. Then to determine the value of NR we deduct the product of AP by the mean quantitative weight from the AP. Or, expressed AP-(AP*QW). (5.6*.2714)=1.52. Then, (5.6-1.52)=4.08.

5. To determine the effectiveness or GLEE< score, these numbers are inserted into the base equation: (5.6-4.08)/(10-4.08)= .2567

The effectiveness score shows that in this hypothetical scenario the effectiveness of the Great Lakes Water Quality Agreement is low at .2567 with respect to fish and wildlife consumption. The process we have described above would be replicated for each of the Objectives which enables comparison across objectives in terms of the effectiveness of GLWQA programs and other measures.

Additionally, as we mentioned previously, it is possible conduct effectiveness evaluations within Objectives as well. This function can be done by evaluating each of the individual programs or policies that are nested under a given Objective.

Another option for conducting intra-Objective evaluation is to make use of one of the other

frameworks we discussed earlier in this report. The WCPA framework, for example, uses a standardized evaluation technique which is particularly well suited to comparing like programs (e.g. effectiveness of sewage treatment facilities). However, it is not as useful for comparing unlike programs (e.g. sewage treatment and invasive species programs). Thus, like the other alternative frameworks presented in this report, this approach has an internal value that might be considered if more specific analysis of particular programs is desired.

CONCLUSIONS

In this report, the authors have provided the Commission with an overview of the effectiveness literature and a review of several relevant effectiveness frameworks in the evaluation literature which can assist the IJC in assessing how effectively the Objectives of the GLWQA are being met. We have recommended a Qualitative-Quantitative Effectiveness Framework (QQEF) based on a case study of the Elbe River as the framework that best provides the breadth and comparative values necessary to encompass the diverse Objectives of the GLWQA.

The authors then developed and described an application of this approach to assessing the effectiveness of Great Lakes programs and other measures in addressing GLWQA Objectives, as well as the relative effectiveness of the programs nested within these Objectives. They further describe a hypothetical situation to demonstrate how a Great Lakes Environmental Effectiveness Metric or GLEEM score can be derived and used to usefully compare the relative effectiveness of GLWQA programs and other measures in

meeting these Objectives.

After GLEEM scores are developed for the nine GLWQA Objectives, the Commission might wish to further assess some unique programs within each objective or improve the implementation of programs found to be somewhat ineffective. Some of the alternative frameworks identified in this report and its appendices might provide additional instruction or approaches/process improvements for the Commission to consider as it continues to GLWQA responsibilities.

A Few Additional Remarks

Whether a program is "effective" in terms of achieving its desired outcome is a different question than what makes a program effective. Our focus has quite clearly been on proposing an evaluation framework that answers the first question and, depending upon how that information may be used, that may be sufficient in and of itself.

We also recognize, however, the value of answering the second question and digging deeper into the operational performance of a particular program- the output issue. The last two models described in this report, in particular, touch upon the latter concept, as they both offer a forensic look at what contributed to decision making, how resources were allocated, the tendencies of leadership, etc.

Brief case summaries are provided immediately below to provide a flavor of the contributions they may make to the effectiveness framework developed in this report. The full cases or articles are provided in the report Appendices.

Case Study Summaries

Oslo-Potsdam Solution: This framework uses mixed methodologies taken from both quantitative and qualitative research techniques. In Dombrowsky's Elbe River study, he made use of available data indicators and expert interviews to derive values for the no-regime counter factual, actual performance, and collective optimum. Those values are inserted into a base equation where Effectiveness (E)=(AP-NR)/(CO-NR). The effectiveness evaluation results in an effectiveness score. This effectiveness evaluation framework is particularly powerful for evaluating unlike program areas such as spill prevention techniques for sewage treatment facilities and invasive species control policies.

WCPA Framework (Level 3): The WCPA framework, developed by the International Union for the Conservation of Nature's World Commission on Protected Areas, identifies three levels of evaluation. The framework proposes a standard approach to evaluating common questions in different arenas. Parks Canada used this model to evaluate a series of questions such as: Is the park losing native species? Are selected indicators within acceptable range? Are herbivores and predators playing their role? Are biological communities at mix of ages and spacing that will support natural biodiversity? The framework directs that indicators and attributes are selected for monitoring across a system (the parks and protected areas under the control of Parks Canada, in this case), and that evaluation against common indices enables the researcher to assess effectiveness using common terminology and standards.

Conservation Excellence Model: The Conservation Excellence Model is built upon the shoulders of total quality management and other enterprise-wide quality and effectiveness

measures used frequently in private industry. As an outcome evaluation model, it arcs the lines of inquiry on goal achievement. The model does not dive deeply into establishing causal links between observed changes in the goal or outcome and the policy or program that is ostensibly the object of the evaluation. We note, however, that there are techniques embedded in other frameworks we have discussed that can be imported to aid in probing the causal link.

Conservation Measures Partnership-Conservation Audit: The conservation audit has been applied as an impact/effectiveness evaluation technique within the adaptive management loop and has been used in numerous location by the Nature Conservancy and other members of the Partnership. There are some good practices that may be embedded in the audit process. For example, the audit process is not punitive and is looking to draw forward practices and policies that have been effective in a given area and to analyze what made it effective. This is a formative evaluation tool, and as such does not render judgment, per se, on effectiveness but can be used in a post effectiveness evaluation context to diagnose the causes of poorly performing practices and standardize practices that are achieving desired results.

References

Black, S. and J. Groombridge (2010). "Use of a Business Excellence Model to Improve Conservation Programs." Conservation Biology **24**(6).

Bronson, J. and B. Noble (2006). "Measuring the effectiveness of Parks Canada's environmental management system: a case study of Riding Mountain National Park " The Canadian Geographer **50**(1).

CMP (2007). Open Standards for the Practice of Conservation Version 2.0, Conservation Measures Partnership.

Dombrowsky, I. (2008). "Institutional design and regime effectiveness in transboundary river management-the Elbe water quality regime." Hydrology and Earth System Sciences **12**.

Helm, C. and D. Sprinz (2000). "Measuring the Effectiveness of International Environmental Regimes " The Journal of Conflict Resolution **44**(5).

Hockings, M., S. Stolton, et al. (2006). Evaluating Effectiveness: A framwork for assessing management effectiveness of protected areas. Best Practice Protected Area Guideline Series P. Valentine.

Hovi, J., D. F. Sprinz, et al. (2003). "The Oslo-Potsdam to Measuring Regime Effectiveness: Critique, Response, and the Road Ahead " Global Environmental Politics **3**(3).

Hovi, J., D. F. Sprinz, et al. (2003). "Regime Effectiveness and the Olso-Potsdam Solution : A Rejoinder to Oran Young " Global Environmental Politics **3**(3).

Johansson, M. V. (2006). "Incentives and Outcomes: Evaluation of a Swedish Environmental Subsidy Programme." Journal of Environmental Planning and Management **50**(3).

Langbein, L. and C. L. Felbinger (2006). Public Program Evaluation A Statistical Guide Armonk, New York M.E. Sharpe

Mickwitz, P. (2003). "A Framework for Evaluating Environmental Policy Instruments." Evaluation **9**(4).

Miles, E. L., A. Underdal, et al. (2002). Environmental Regime Effectiveness: Confronting Theory with Evidence Cambridge, MA, The MIT Press.

Owen, J. M. and P. Rogers (1999). Program Evaluation: Forms and Approaches, Sage

Parrish, J. D., D. P. Braun, et al. (2003). "Are We Conserving What We Say We Are? Measuring Ecological Integrity within Protected Areas." BioScience **53**(9).

Scriven, M. (1991). Beyond formative and summative evaluation. Evaluation and education at quarter century. G. W. Mclaughlin and D. C. Phillips. Chicago, University of Chicago Press.

Sprinz, D. (2002). "Research on the Effectiveness of International Environmental Regimes: A Review of the State of the Art". Final Conference of the EU Concerted Action on Regime Effectiveness (IDEC), 09-12

Underdal, A. (1992). "The concept of regime "effectiveness"." Cooperation and Conflict **27**(3).

Vollenweider,J. (2012) "The Effectiveness of International Environmental Agreements." Center for International and Comparative Studies. Zurich.

Vedung, E. (2000). Public Policy and Program Evaluation New Brunswick, New Jersey

Transaction Publishers.

Wholey, J. S., H. P. Hatry, et al., Eds. (1994). Handbook of Practical Program Evaluation San Fransisco Jossey-Bass Publishers

Young, O. R. (2001). "Inferences and Idices: Evaluating the Effectiveness of International Environmental Regimes." Global Environmental Politics **1**(1).

Young, O.R. (2011). "Effectiveness of international environmental regimes: Existing knowledge, cutting-edge themes, and research strategies". Proceedings of the National Academy of Sciences 108(50).