Oil Sands
Regional Aquatic Monitoring Program (RAMP)
Scientific Peer Review of the Five Year Report (1997-2001)

Submitted to:
RAMP Steering Committee
February 13, 2004

Prepared by:
G. Burton Ayles, Winnipeg, Manitoba
Monique Dubé, Saskatoon, Saskatchewan
David Rosenberg, Winnipeg, Manitoba

TABLE OF CONTENTS

EXECUTIVE SUMMARY	**iii**
INTRODUCTION	**1**
METHODOLOGY FOR THE REVIEW OF RAMP	**3**
ASSESSMENT OF CLIMATE AND HYDROLOGY COMPONENT	**8**
ASSESSMENT OF WATER QUALITY COMPONENT	**24**
ASSESSMENT OF SEDIMENT QUALITY COMPONENT	**32**
ASSESSMENT OF BENTHIC INVERTEBRATES COMPONENT	**35**
ASSESSMENT OF FISH POPULATIONS COMPONENT	**40**
ASSESSMENT OF AQUATIC VEGETATION COMPONENT	**44**
ASSESSMENT OF ACID SENSITIVE LAKES COMPONENT	**49**
OVERALL ASSESSMENT OF THE RAMP AND RECOMMENDATIONS FOR THE FUTURE	**58**
ACKNOWLEDGEMENTS	**67**
BIBLIOGRAPHY	**68**
APPENDIX I. RAMP OBJECTIVES	**72**
APPENDIX II. REVIEWERS BIOGRAPHIES	**73**
APPENDIX III. OIL SANDS REGIONAL AQUATIC MONITORING PROGRAM (RAMP) SCIENTIFIC PEER REVIEW OF THE FIVE YEAR REPORT (1997-2001): REVIEW OF BIOSTATISTICS	**81**
APPENDIX IV. OIL SANDS REGIONAL AQUATIC MONITORING PROGRAM (RAMP): SCIENTIFIC PEER REVIEW OF THE FIVE YEAR REPORT (1997-2001): REVIEWS OF INDIVIDUAL RAMP COMPONENTS. TABLE OF CONTENTS	**116**

EXECUTIVE SUMMARY

The Oil Sands Regional Aquatics Monitoring Program (RAMP) is a multi-stakeholder, multi-objective long-term program, designed to incorporate both traditional and scientific knowledge to address monitoring needs in the Region. The RAMP organizational structure is a Steering Committee with representatives from the oil and gas industry, other industries in the Region, First Nations, and provincial and federal governments. In 2003, the Steering Committee initiated an independent scientific peer review of the monitoring program to ensure that the program continued to meet monitoring objectives and to ensure that knowledge and understanding being applied was appropriate to the task.

The specific purposes of the review were:

1. to assess the program for adequacy against the relevant objectives of RAMP;
2. to evaluate the program design, methods, and results of RAMP with respect to the objectives of detecting change, determining regional variability and cumulative effects, and verifying EIA predictions; and
3. to provide recommendations for changes to the program, including an assessment of the potential impact of those changes to the integrity of the program in the future.

The review was based primarily on a Five Year Report that presented the results of the monitoring program between 1997 and 2001. Discipline specialists carried out reviews of the various components of RAMP viz., climate and hydrology, water quality, sediment quality, benthic invertebrates, fish populations, aquatic vegetation and acid sensitive lakes. The review has been structured around the three fundamental goals of RAMP as identified in the Five Year Report, i.e. characterizing existing variability, detecting regional trends and cumulative effects, and monitoring to verify EIA predictions.

Narrative summaries of the assessments, including major gaps and recommendations for each component, are presented in this report. As well, there is a general assessment of common themes or issues and recommendations for future improvement of the overall program. Details of the assessments of the various components or programs of RAMP are found in Appendix IV of this report.

The reviewers found many signs of positive progress with RAMP. The very existence of a major regional aquatic monitoring program is a very positive sign for Alberta. Beginning joint monitoring by companies in 1997 was a progressive initiative leading to benefits now and in the future. The companies involved are to be commended for their vision and their significant financial contribution over the years. A long-term initiative such as this is rare. As well, the RAMP initiative to draw individual components into a comprehensive regional aquatic monitoring program is seen as a positive step towards relevance and effectiveness. This program offers a significant opportunity to ensure environmental protection, support environmental rehabilitation in the future and enhance our level of knowledge and understanding of boreal aquatic ecosystems in a disturbed and undisturbed setting.

The general consensus of the reviewers was that the Five Year Report was well organized and written in a manner that is accessible to most stakeholders, with a few exceptions. It fairly describes the evolution of RAMP over the years and, with the unfortunate exception of the aquatic vegetation and the acid sensitive lakes programs, which were not addressed, it is a good description of what was done. The problems with the report are found in lack of details of methods, failure to describe rationales for program changes, examples of inappropriate statistical analysis, and unsupported conclusions.

That being said, the reviewers raised significant concerns about the Program itself. They felt there was a serious problem related to scientific leadership, that individual components of the plan seemed to be designed, operated and analyzed independent of other components, that there was no overall regional plan, that clear questions were not been addressed in the monitoring and that there were significant shortfalls with respect to statistical design of the individual components.

Based on the results of the individual reviews the Design and Integration Team presents the following recommendations for future improvement of monitoring of the aquatic environment of the Oil Sands Region of Alberta:

I. Organizational Recommendation on Scientific Leadership

We recommend that RAMP establish a new independent position of project scientific leader reporting to the RAMP Steering Committee and responsible for the overall scientific design of the program and ensuring program quality and relevance through independent peer review. RAMP should also establish an ongoing system of independent scientific input to the program through (1) informal or formal commentary on early ideas and initial plans; (2) workshops and planning sessions that involve independent researchers, and RAMP contractor staff and RAMP technical committee members in interchange and debate; (3) formal written review of monitoring plans; and (4) formal review of progress on a periodic basis.

II. Primary Technical Recommendations

1. Adoption of an Ecosystem Approach and Decision-Making Strategy. We recommend that RAMP adopt a strategic, integrated, regional monitoring design and decision-making strategy for measurement of development-related change at an ecosystem level while incorporating site-specific needs. Monitoring must fit within the context of an adaptive management framework and focus beyond project-specific needs. This approach should:

- Consider how decisions on change will be made and the information that is required to make those decisions. For example, what indicators will be measured to assess a particular development activity? What will the indicator be compared against to determine when a change has occurred? Will changes of a certain magnitude and direction trigger a specific line of decisions or an approach to greater monitoring intensity? What will the process be if water

quality indicators show a change but no change was measured in fish indicators?

- Consider the development projections to 2020 in the oil sands area and select strategic monitoring locations accordingly. Depending upon the watershed, development level, and physical, chemical, and biological characteristics the monitoring approach can be customized. Sampling intensity and frequency can also be customized;
- Integrate RAMP components (i.e. hydrology, water and sediment quality, benthic invertebrate community structure, fish population health, aquatic vegetation and acid sensitive lakes) at integrated monitoring stations;
- Use adaptive feedback loops within and among components for constant examination of experimental designs and results; changes should be made to the program based on solid results rather than on speculation;
- Show clear links to objectives and have clearly stated hypotheses or testable study objectives; and
- Ensure that all terms, especially statistical ones, are defined and used precisely in reports, and a glossary for all component subject areas be produced as an aid to authors and readers of reports. Precise use of terms aids understanding.

2. Adoption of Effects-Based Monitoring within the Strategy. We recommend that RAMP orient its efforts towards effects-based monitoring. The objective should be to document environmental change occurring as a result of development, not to carry out descriptive studies. Included in the effects-based approach should be the following:

- Selection of key response indicators for each RAMP component, based upon potential changes resulting from oil sands development;
- On-going synthesis of information related to development pressures including type of development activity, location of activity, stressors released, effects predicted, assumptions used in predictive tools, location of modeling nodes, etc. A monitoring program designed to monitor development-related change cannot do so in the absence of information on the development. This was recognized as a significant shortfall of the RAMP program. Reviewers recognized that much of this information is likely included in the EIA reports. However, effects-based monitoring mandates an on-going comparison between development activities and environmental condition. One without the other will not measure development-related change;
- Establishing a core level of consistency for sample station selection, indicator selection, sampling frequency and timing that does not change from year to year;
- Selection of reference and “low-impact” stations within or outside the Region for each component subject area. Those subject areas that can go into an established biomonitoring program will get this benefit automatically;
- Use of biostatistical analyses that report statistical confidence levels and power analyses for indicators of change. These statistical results are critical to assist with interpretation of the environmental changes to establish confidence in the decision-making strategy;

- Consideration of the knowledge and understanding gained from other successful effects-based monitoring programs that measure development-related change relative to natural variability; for pertinent subject areas such as water quality, benthos, fish and possibly aquatic vegetation, a bona fide, regional biomonitoring program (Environmental Effects Monitoring [EEM] or the Reference Condition Approach [RCA]) should be initiated; and
- Incorporation of other existing regional information such as NRBS, NREI, PERD, EEM, the Muskeg River design initiative (CEMA) and information collected independently by industry. Future periodic summary reports, such as the next Five Year Report, should incorporate monitoring results and studies from programs other than RAMP, if the information contributes to the objectives.

3. Testing Environmental Impact Assessments (EIA) Prediction. We recommend that RAMP complete an exercise to test predictions from already completed EIAs using actual data generated on a site or sites. As a first step in this evaluation, RAMP should prepare a synthesis or summary, on a project-specific basis, of what the impact predictions were for different project activities, including location and timing of impact and Valued Ecosystem Components (VECs) affected.

4. Development of an Information Management System. We recommend that RAMP establish a comprehensive information management and assessment system, including an electronic database management system that would enable electronic reporting of raw data in a standard and consistent format, interchange of data among component subject areas, and on-going assessment of data using consistent analyses.

5. Increased Emphasis on the Athabasca River as a Priority Watershed. We recommend that RAMP use the Athabasca River as a central focus for monitoring across component subject areas because it is the largest and most important aquatic ecosystem in the region and the natural recipient of the effects of oil sands development.

II. Secondary Technical Recommendations

1. Contributions to New Knowledge. We recommend that RAMP recognize the importance of creating new knowledge and incorporating this knowledge into the monitoring program through an adaptive management framework.

2. Traditional Ecological Knowledge (TEK). We recommend that RAMP actively promote the use of TEK by incorporating it into the design of scientific programs. Key indicators for future monitoring and the interpretation of results need to be identified, and specific, ongoing programs should be devoted to observing changes in these key indicators.

3. Publications. We recommend that RAMP initiate a policy of encouraging individuals and the contractor to publish monitoring data and new knowledge in

established technical and primary publications as well as in-house reports. RAMP should also establish a RAMP Technical Report Series for wider distribution of monitoring results within the region, provincially and nationally.

INTRODUCTION

Environmental Impact Assessments (EIAs) in the Oil Sands Region of northeastern Alberta document baseline environmental conditions and predict effects of proposed developments. Understanding long-term natural variability in the region is essential in determining if changes to the aquatic environment are due to the effects of development, natural extremes or both. The Oil Sands Regional Aquatics Monitoring Program (RAMP) is a multi-stakeholder, multi-objective long-term program, designed to incorporate both traditional and scientific knowledge to address the monitoring needs in the Region. The RAMP organizational structure is a Steering Committee with representatives from the oil and gas industry, other industries in the Region, First Nations, and provincial and federal governments. There are four subcommittees: Technical, Communications, Finance and Investigators. The RAMP technical program was initiated in 1997 and annual reports have been produced since 1997 (RAMP, 1997-2001). In May 2003, under the direction of the RAMP Steering Committee, the contractor, Golder and Associates, (Golder)[1] completed a Five Year Report covering the period 1997 to 2001 (RAMP, 2003).

In 2003, the Steering Committee initiated an independent scientific peer review of the monitoring program to ensure that the program continued to meet monitoring objectives and to ensure that knowledge and understanding being applied was appropriate to the task. The Steering Committee established a Review Team[2], who contracted with us, Dr. Burton Ayles, Dr. David Rosenberg and Dr. Monique Dubé (the Design and Integration Team), to design, lead and coordinate the review.

The specific purposes of the review were:

1. to assess the program for adequacy against the relevant objectives of RAMP;
2. to evaluate the program design, methods and results of RAMP with respect to the objectives of detecting change, determining regional variability and cumulative effects, and verifying EIA predictions; and
3. too provide recommendations for changes to the program including an assessment of the potential impact of those changes to the integrity of the program in the future.

This report presents the results of our review. Our description of the structure of the review includes a brief discussion of the need for an independent scientific peer review in environmental and resource management programs, the contribution such reviews can make to improved planning and decision making and considerations of what should be included in a review. We then describe the specific process for this review. The subsequent sections are summaries of the independent scientific peer reviews of the various components of RAMP viz., climate and hydrology, water quality, sediment quality, benthic invertebrates, fish populations, aquatic vegetation and acid sensitive lakes. These assessments include specific recommendations for each component. The

[1] Beginning in 2003, RAMP contracted with Hatfield Consulting (Hatfield) to carry out the monitoring aspects of RAMP.
[2] Bryan Kemper, Christine Brown, Preston McEachern, and Mark Spafford.

final section provides our overall assessment and recommendations. A separate Appendix (Appendix IV) contains detailed reviews of each of the components following a prescribed template.

METHODOLOGY FOR THE REVIEW OF RAMP

General Approach

Peer review has traditionally been used to assure the quality of research carried out in government and academic laboratories. The process for large-scale planning and evaluation of complex environmental and resource management programs is neither well established nor universally accepted (Fleishman, 2001), but it is our belief that independent scientific peer review can contribute significantly to the design, operation and modification of monitoring of the aquatic environment of the Oil Sands Region. Independent scientific review can help to ensure that decisions reflect the best current scientific knowledge. It can help to focus RAMP on objective scientific variables apart from historical, economic or social variables. It will help to raise the trust of all stakeholders in the RAMP. And, perhaps most importantly, without an independent scientific peer review, any claims of objective and scientific validity may be suspect (adapted from Meffe et al., 1998). It was our intention that this review address only science issues in the belief that this approach should help RAMP to clarify areas of potential overlap between science and non-science. Non-science issues fall under the purview of RAMP, the companies and the stakeholders, not our review.

Our independent scientific review should be a tool for improvement of RAMP. Most environmental regulatory agencies accept the "precautionary principle" as a general guideline for doing business, and many environmental regulators use terms like "current best practices" or "best available science" (Dorcey and Hall, 1981; CSTA, 1999; Bisbal, 2002; WSOCD, 2002) and the Alberta Energy Utilities Board uses the phrase "best available technologies". Our expectations are that RAMP programs should follow the precautionary principle and they should be using the best science. As well as an expectation of good science, there are other characteristics that guided us in the development of the plan for this review. We felt it was important that bias and special interest were minimized so we selected reviewers who are not involved directly in aquatic monitoring or research in the Region. This choice made the task somewhat harder because individual reviewers did not have much background knowledge and had to take time to familiarize themselves with the overall area. The reviewers possibly being unaware of some pertinent information or program introduced the possibility of errors in our assessments and recommendations. However, we feel that the risks of error were justified by the benefits of impartiality. We tried to minimize such mistakes, and ensure that all relevant information was considered, by reviewing additional information beyond the Five Year Review Report (see template reports for details), holding an interim meeting with the RAMP Technical Committee and periodic exchanges with the Review Team. We have attempted to ensure that our conclusions and recommendations are consistent with available scientific information and that our assumptions are explicit. We did not enter into the planning for this review with a standardized or set format in mind. Circumstances for independent scientific review of large-scale environmental projects vary greatly by issue, and we felt that our process had to be tailored specifically to the monitoring program at hand. We developed our process based on our personal knowledge of research management and control, reviews of current literature on independent scientific reviews of natural resource management projects (e.g. Meffe et. al.

1998; CSTA 1999; Roosenberg, 2000; Fleishman 2001). Our initial plans were modified after discussion with the Review Team and individual reviewers, and we modified the process as we proceeded and problems were identified. We did not always agree with suggestions from the individual reviewers or the RAMP Review Team and, ultimately, the responsibility for the review process is ours. We know that there are risks associated with alternative monitoring plans and we hope to give RAMP the knowledge to make the necessary decisions.

Scope of this Review

Our review is based on the published annual and summary reports of RAMP. Specific programs in RAMP were established each year by committees and subcommittees after consultation with industrial, aboriginal, environmental and regulatory stakeholders and expert independent consultants. As the Oil Sands Region experienced rapid growth from 1997 to 2001, changes to RAMP were made annually. These changes not only affected RAMP's objectives, and organizational structure, but the study area and study design as well. Potential sampling methods, sentinel species and reference lakes and streams were also evaluated during this period. Some methods were adopted and then abandoned during following years. Through the years, RAMP included the following environmental monitoring in the Oil Sands Region:

- hydrology and climate (monitoring began in 1995, but became a component of RAMP in 2000);
- water quality in rivers (1997 to 2001);
- sediment quality in rivers (1997 to 2001) and wetlands (1998 to 2001);
- benthic invertebrates in rivers (1997, 1998, 2000 and 2001) and two lakes (1997 to 2001);
- fish populations in rivers (1997 to 2001);
- aquatic vegetation (1999 to 2001); and
- acid sensitive lakes (1999 to 2001).

Each year the detailed monitoring activities and results for that particular year were summarized in an annual report prepared by Golder. The Five Year Report, completed in May 2003, includes the analysis of data over the five year period from 1997 to 2001, where Golder considered sufficient data were available viz.: climate and hydrology, water quality, sediment quality, benthic invertebrates, and fish populations. Components of the RAMP program that did not have sufficient data, such as the aquatic vegetation and acid sensitive lakes, were not included in the Five Year Report.

Although there are eight stated goals for RAMP (see RAMP Terms of Reference for details) the Five Year Report addresses only the three program objectives considered most relevant to aquatic monitoring (Appendix I). They are:

1. Characterizing Existing Variability - To collect scientifically defensible baseline and historical data to characterize variability in the oil sands area (the capacity to detect change was of particular importance for reviewers to consider).
2. Detecting Trends and Cumulative Effects - To monitor aquatic environments in the Oil Sands Region to detect and assess cumulative effects and regional trends

(the capacity to detect cumulative effects and trends for new disturbances was of particular importance for reviewers to consider).

3. Monitoring to Verify EIA Predictions - To collect data against which predictions contained in EIAs can be verified;

This peer review focused primarily on the Five Year Report but incorporated acid sensitive lakes and aquatic vegetation to the extent that they appear in the annual reports. Our analyses of the individual components of the RAMP summarized in the following chapters and detailed in Appendix IV has been structured around the three fundamental goals of RAMP as identified in the Five Year Report. Other RAMP goals were addressed as they were deemed relevant by the reviewers.

Process for the Review

Reviewers - The Review Team established a number of task areas and tentatively identified a number of possible reviewers for specific areas. We contacted the independent scientists and biologists who would carry out the individual reviews. Of our 16 primary reviewers (Table 1 and Appendix II) four were from universities, seven from government agencies and five were consultants. The Review Team approved all of the reviewers.

Table 1. RAMP independent scientific peer review. Names, institutions and components reviewed by individual reviewers.

Reviewer and Institution or Agency	**Component Reviewed**
Neil Arnason, Ph.D., University of Manitoba,	Biostatistics
Burton Ayles, Ph.D., Consultant	Coordinator
Jan Barica, Ph.D., Consultant (UNEP)	Water quality and acid sensitive lakes
Brian Brownlee, Ph.D., Environment Canada	Sediment quality
Uwe Borgman, Ph.D., Environment Canada	Sediment quality
Martin Carver, Ph.D., Consultant	Climate and hydrology
Monique Dubé, Ph.D., Environment Canada	Benthic invertebrates, water quality, fish populations, Coordinator
Nancy Glozier, M.Sc., Environment Canada	Water quality
Kelly Munkittrick, Ph.D., University of New Brunswick	Fish populations
John Post, Ph.D., University of Calgary	Fish populations
David Rosenberg, Ph.D., Consultant	Benthic invertebrates, Coordinator
Carl Schwarz, Ph.D., Simon Fraser University	Biostatistics
Brian Souter, M.Sc., Department of Fisheries and Oceans	Fish abnormalities

Stephanie Sylvestre, M.Sc., Environment Canada	Benthic invertebrates
Alan Thomson, M.Sc., Consultant	Climate and hydrology
Michael Turner, Ph.D., Department of Fisheries and Oceans	Acid sensitive lakes
Marley Waiser, Ph.D. Environment Canada	Aquatic vegetation

Components - RAMP programs were reviewed by teams of two to three specialists in a particular subject or component viz., climate and hydrology, water quality, sediment quality, benthic invertebrates, fish populations, aquatic vegetation and acid sensitive lakes. The format used for the assessment of each program was generally similar but each team had the latitude to address issues more specific to its area. Biostatistics were addressed within each component and a separate biostatistics report was prepared as well (Appendix III). A biostatistics specialist was available for consultation/questions from individual teams and review final template reports and narrative summaries for appropriateness of statistical recommendations.

Template for Review – Because of the complexity of the program and of a review involving so many individuals, we felt it necessary to give the reviewers significantly more guidance than they would have received if asked to review a scientific paper for a journal publication. Reviewers were asked to report on inadequacies in the report(s) that could be corrected through reanalysis/reinterpretation of data or results and reported in future annual or periodic reports. They were also asked to report on inadequacies that required changes to the program itself. A template was provided for the individual program reviews. This template contained the elements that the Design and Integration Team considered desirable in a well-designed regional aquatic monitoring program. The specialist teams were asked to prepare separate template reports for each of the three primary RAMP objectives that the Five Year Report had been structured around. The template was relatively comprehensive and but all of the points would necessarily relate to each of the objectives. Common elements considered in the review of each objective for each subject area were: assessments of relevance to objectives; appropriateness of experimental design; interpretation of results and conclusions; nature of outputs; linkages to other components and programs; an assessment of gaps, omissions and recommendations; and an assessment of the proposed program for 2003 to 2009. Specific questions or directions to be considered were provided for each of the elements (see Appendix IV). A draft template was circulated to reviewers and members of the Review Team for comments before beginning the review but we remain responsible for any weaknesses, or strengths, of the approach. However, some elements that we thought should be included in a review (e.g. cost effectiveness), could not be addressed because the necessary information was not available.

Narrative Reports - Based on the detailed template reports the specialist teams were asked to prepare summaries for their components following a common format. They were asked to describe the scope of their review and any special perspectives that they might have. The assessment section focused on the adequacy of the monitoring and

whether it met the objectives. Reviewers were asked to emphasize the most important/relevant aspects and leave secondary aspects to the template reports (Appendix IV). They were asked to make recommendations that addressed gaps identified in their review, and to provide enough details on implementation to provide guidance for further program design. The narrative summaries formed the basis of the assessments and recommendations in this report. With the exception of formatting, we did not edit these summaries. They are the creation of the specialist teams.

Integration – The various elements of an aquatic monitoring program need to be carefully integrated if the program is to be effective, and the same approach applies to any review of that program. We addressed this need for integration with ongoing interchange between the specialist teams and the Design and Integration Team, cross appointments of some of the specialists to more than one component and review of each of the component reports by the Design and Integration Team. The Design and Integration Team was also responsible for identifying common issues, discussing them with the Review Team and the RAMP Technical sub-committee during an interim meeting, and preparing the overall assessment and recommendations.

Recommendations – Recommendations from the specialist review teams will be found in the narrative sections that follow. More detailed comments are in the template assessments, particularly in the final section under each objective. The overall recommendations for this review were prepared by the Design and Integration Team. We reviewed all of the component reports looking for common themes in the assessments and recommendations. Our overall recommendations cross RAMP component areas and RAMP objectives. They are in general order of priority and they are closely inter-related.

ASSESSMENT OF CLIMATE AND HYDROLOGY COMPONENT
PREPARED BY ALAN THOMSON AND MARTIN CARVER

1.0 Introduction

The climate and hydrology section of the RAMP Five Year Report is reviewed in this section. The review was conducted by two professional hydrologists with wide experience in the assessment of regional hydrological monitoring programs (see Appendix II). The review follows the format of the RAMP Five Year Report and is divided into three sections, each addressing one of the first three principle RAMP objectives:

1. Charaterizing existing variability;
2. Detecting and assessing cumulative effects; and
3. Monitoring to evaluate environmental impact assessment (EIA) predictions.

This review comments on the design, methods, results and conclusions of the RAMP climate and hydrology section. The review also makes an overall assessment of the section, comments on the gaps in the section and makes recommendations for changes to the overall climate and hydrology monitoring program, data analysis and data reporting. Lastly, comments are made as to the future direction of the climate and hydrology program within RAMP.

This review is a condensed version of the review found in Appendix IV, and includes only the more important issues. The reader should refer to the review in the appendices for additional discussion, analysis and recommendations concerning the climate and hydrology section.

2.0 Charaterizing Existing Variability

Under the title of the first primary objective, Charaterizing existing variability, the report discusses the location of long-term climate and hydrometric stations installed by Environment Canada to monitor both climate and hydrologic parameters in watersheds in the oil sands development region. Data retrieved from the climate and hydrometric stations are analyzed for variability in temperature, precipitation, water yield, flood discharge, and low-flow discharge. Watershed response to precipitation input is discussed and comparisons are made between watersheds. Recommendations are made in the Five Year Report that will increase both the quantity and quality of data collected. This primary objective is further described as “collecting scientifically defensible baseline and historical data to characterize variability in the oil sands area” (p. 2-1 of the Five Year Report). The technical subcommittee established more detailed objectives, which are:

- to characterize the natural variation in climatic and hydrologic parameters, including precipitation, air temperature, water yield, flood peak discharges and low flows in the Oil Sands Region and identify linkages between climatic and hydrologic parameters; and
- to define baseline ranges for climatic and hydrologic parameters for the area monitored by RAMP.

The reviewers assessed this objective for the adequacy of the monitoring design, the methods used, the statistical analysis and the results and conclusions reached. Recommendations are also made under each of these subjects.

2.1 Monitoring Design

The reviewers had three major concerns about the monitoring design as outlined in the report:

- monitoring program design;
- monitoring groundwater; and
- monitoring of low flows.

2.1.1 Monitoring Program Design

To assess whether the monitoring program design is adequate to characterize natural variation in climate and hydrologic parameters, it is necessary to understand the rationale for the design of the monitoring program. The few details provided for the long-term hydrometric stations are mostly location oriented. There is little discussion of each station's attributes, the quality of data, data limitations, and most importantly, what role the station plays in the overall monitoring program. As an example, the station location rationale that exists for Station S4A (a short-term hydrometric station) as presented in Table 2.2 represents the absolute minimum in detail that the reviewers consider adequate. The apparent lack of monitoring station rationale documentation points to the need for a monitoring design and analytical plan that would outline RAMP's climate and hydrology monitoring objectives, strategies, and background information. This information would help in developing a scientifically defensible implementation, data collection, and reporting plan. See additional discussion under Objectives 2 and 3.

It is unclear from reviewing the Five Year Report whether the monitoring program is able to describe "the natural variation in climate and hydrologic parameters…in the Oil Sands Area". The focus study area, as defined by Figure 1.2 in the Five Year Report, is much larger than the area covered by the many long- and short-term hydrometric and climate monitoring stations. In addition, it is not clear from the Five Year Report the scope of the monitoring required and the level of detail and certainty required. Should all ecoregions be monitored? Should all geomorphologically distinct zones be represented? Should a complete range of watershed areas be monitored? (For example, there is no watershed represented in the 2,000-5,000 km^2 category.)

Due to the issues raised above, it is difficult to assess whether the current monitoring network is sufficient to characterize natural variation in climate and hydrologic parameters in the Oil Sands Region. It is recommended that a detailed discussion be presented in subsequent monitoring reports that outlines the monitoring design principles, objectives, and how the current long-term monitoring network addresses those objectives.

2.1.2 Monitoring Groundwater

Local groundwater resources are predicted to be significantly affected by oil sands developments as presented in recent EIAs for CNRL's Horizon and Shell's Jackpine Phase 1 projects. As such, oil sands developers are required to monitor local

groundwater resources. Local and regional groundwater resources, however, are not monitored by RAMP. Groundwater contribution to baseflows predominate over all other water sources during low-flow winter months. Since the overwintering survival of many aquatic species is dependent upon the quality and quantity of baseflows, the reviewers consider it important to monitor local and regional groundwater resources. It is recommended that the RAMP hydrology monitoring program consider:

- monitoring local and regional groundwater by data acquisition from oil sands developers; and
- groundwater monitoring station placement in areas considered environmentally sensitive that are outside of oil sands developers' existing and proposed groundwater monitoring zones.

2.1.3 Monitoring of Low Flows

The collection of low-flow data seems to be a weak link in the hydrometric data collection system. For six of the seven long-term monitoring stations, sampled data do not exist post-1987 for the November-to-February period when low flows predominate. It is imperative that low flows be monitored at all hydrometric stations as low-flow discharge often dictates survival of aquatic biota, including fish, during winter months. The reviewers acknowledge that the Five Year Report recommends increased winter monitoring at some stations. However, the report notes that some small streams and tributaries may freeze to the bottom and thus provide no baseflow, and therefore do not warrant monitoring. Until this assumption can be proven over time and through various wet and dry cycles, 12-month monitoring should be conducted at all hydrometric stations. It is recommended that discharges be monitored from November through February for all hydrometric stations, regardless of known or assumed historical or current discharge characteristics.

2.2 Methods

The methods used to analyze the data appear appropriate. However, there are some issues of concern that should be addressed in future reports. These issues are detailed in the review found in Appendix IV.

2.3 Statistical Analysis

A standard statistical analysis was performed on both the long-term climate and hydrometric data. Mean annual discharge, frequency analysis, range of data, coefficient of variation, standard deviation and skewness were calculated for water yield, flood discharge and low-flow discharge. The report also computed the correlation between precipitation and water yield for each of the seven long-term hydrometric stations, and commented on the variability of hydrologic parameters between monitored watersheds. Concerns about the statistical methods used focus on two areas:

- time period for frequency analysis of hydrometric data; and
- frequency analysis using interpolated data.

2.3.1 Time Period for Frequency Analysis of Hydrometric Data

Frequency analysis of hydrometric data requires a significantly long period of record for the results to have a reasonable degree of statistical confidence. The report notes this and

adds that "…48 years of record would be required to provide a 100-year flood estimate accurate to within 25% of the expected population value at a 95% confidence limit…" (p. 3-11). The data used to conduct a frequency analysis, with the exception of the Athabasca River station number 07DA001, fall short of recommended record periods to produce statistically meaningful results. Of the seven hydrometric stations in the region chosen for long-term variability analysis, one has 44 years of continuous data, and the remaining six have, on average, 28 years of continuous and interpolated data. It is possible to conduct frequency analysis on such data, but it is recommended that the statistical confidence level be reported and illustrated using confidence bands on all figures that indicate extreme events. There is concern that the length of data record for six of the seven hydrometric stations is insufficient to accurately predict future flood events with acceptable statistical confidence at this time.

2.3.2 Frequency Analysis Using Interpolated Data

Since 1987, six of the seven long-term hydrometric stations have not sampled discharges for the period November to February, representing 33% of the year. In order to conduct a complete analysis, reasonable methods are employed to fill the annual November to February data gap. For some calculations, such as water yield, the error associated with this data interpolation is likely minimal. However, with other calculations, such as low-flow frequency analysis, the statistical error may be large and unacceptable. The confidence level associated with the low-flow analysis is not reported. There is also concern that the interpolated values may be biased data for this analysis since data trends from pre-1987 are used to fabricate the post-1987 data. Although the reviewers recognize that the data gaps must be filled, it is recommended that a detailed discussion of the data-biasing implications and statistical error associated with low-flow data interpolation and analysis be presented.

2.4 Results

The results of the analysis are adequately presented, although the accuracy of some of the results is of concern, given the above comments. There is also a good general description of each watershed's geomorphology and the hydrologic response to precipitation that generally can be expected. The variability on an annual basis of both climate and hydrometric data is also reported satisfactorily. However, what is absent in the report is a detailed variability analysis on other periods; for example, among all of the June records or all of the July records. There is need to report figures and statistics such as variability between months, during periods important to fish and other aquatic biota, between peak flood dates, between low-flow events, during "wet" and "dry" cycles, etc. There are many variability statistics useful in charaterizing watersheds pertinent to biological processes that are not included in this report. It is recommended that time periods and parameters of interest for variability analysis be identified and a thorough analysis be completed and reported. The lack of depth in the variability analysis is the weakest aspect of the reporting for Objective 1.

3.0 Detecting And Assessing Cumulative Effects

The program assessed in this section is addressed to the second primary objective, detecting and assessing cumulative effects, for the climate and hydrology section. The

report examines long-term climate and discharge data from Environment Canada for watersheds in the oil sands development region near Fort McMurray. Trends in precipitation, air temperature, water yield, flood discharge, and low flows are analyzed to identify temporal and spatial patterns in the existing data. Findings are presented and some suggestions put forth to explain the observed trends. The program's ability to detect change is also discussed.

The second primary objective, detecting and assessing cumulative effects, involves monitoring aquatic environments in the Oil Sands Area to detect and assess cumulative effects and regional trends. The technical subcommittee established more detailed objectives:

- to investigate trends over time in precipitation, temperature, water yield, flood peak discharges and low flows, based on available long-term climatic and hydrologic data; and
- to evaluate whether cumulative effects can be evaluated at this time and whether the data collected by RAMP will be appropriate to do so in the future.

The specific reframed objectives sharpen the primary objective by providing clearer statements of what will be addressed under the topic of "assessment and detection of cumulative effects and regional trends". However, the reframed objectives limit the scope of this overall section:

- only precipitation, air temperature, water yield, flood peak discharges and low flows are considered; and
- the assessment of cumulative effects is limited to an assessment of whether they can be evaluated now and in the future.

These changes reduce the scope of Objective 2. Although the following review comments acknowledge and make use of the information provided in these more specific objectives, the review assumes that the broader objective is also to be met, i.e. assessing and detecting cumulative effects and regional trends. The reviewers assessed this Objective for the adequacy of the monitoring design, the methods used, the statistical analysis and the results and conclusions reached. Comments and recommendations that fall under each of these subjects are found below.

3.1 Monitoring Design

Concerns about the monitoring design for the second objective focus on four subjects:

- Analytical plan;
- Strategies used to detect cumulative effects;
- Parameters chosen for analysis; and
- Extent of appropriate data.

3.1.1 Analytical Plan

It is difficult to identify the fundamentals of the analytical plan on which the cumulative effects analysis is built. Cumulative effects studies remain an emerging area of EIA. Given that there has been a lack of consensus on what a cumulative effects analysis should include (e.g. Reid 1993), studies with cumulative effects objectives need to be

clear on what is under consideration. For example, under RAMP what is a cumulative effect? What are the key impacts and potential interactions? How does the monitoring design allow effect to be connected to cause given the measurables and data sets involved both now and in the future? The Five Year Report (p. 2-2) implies a definition of cumulative effects to be "the sum of all the effects on the aquatic environment" and that cumulative effects "are the result of both natural and man-made changes". This definition is confusing because one would think that a study like this would seek to detect whether human-caused impacts are occurring and, if so, what is their total effect? The distinction between human-caused and naturally caused variability should be central to a cumulative effects analysis. The Five Year Report also says (p. 2-2) that the concepts of cumulative effects and regional trends have been "combined" and that "a regional trend, particularly a trend at a downstream location, incorporates the cumulative effect". Again, the idea of combining trends and cumulative effects is confusing and its rationale unclear. Given the centrality of this discussion to the entire objective, it is recommended that the analytical plan for the cumulative effects analysis be worked out and clearly presented. In support of this discussion, it is recommended that the report include a glossary providing a clear expression of the meaning of these and other terms; this would also support greater comprehension by other, less technical readers.

3.1.2 Strategies Used to Detect Cumulative Effects

Beyond the fundamentals, what are RAMP's analytical strategies to detect human-caused impacts and distinguish them from natural variability? Given the physical size of the study area and the scope for development-related impacts superimposed and/or interacting with natural variability, it is to be expected that detection of impact will be a major challenge. Meeting this challenge may require a variety of monitoring approaches and use of alternative data sources, particularly given the large range in applicable spatial and temporal scales and the degree of variability from natural disturbances. RAMP emphasizes the application of before-and-after monitoring both in terms of shorter-term environmental impact assessments and the longer-term cumulative effects assessment. The cumulative impact monitoring relies heavily on eight long-term Environment Canada data sets to establish background trends, yet these may be insufficient for the task (see 3.1.4 below).

What other approaches can be explored? At a minimum, it is recommended that control watersheds be established to act as benchmark comparisons as the oil sands developments are further implemented. It seems that none of the study basins is being held as a control given that all of the study basins shown in Figure 3-3 are within areas that have been or will be developed during the course of RAMP. Depending on the analytical framework, it may be necessary to include a suite of approaches to address data limitations—paired watersheds, analysis of lake bed cores, interpretation of historic airphotos, etc. Opportunistic use of basin "nesting" in the network layout may be helpful in achieving more insights with limited resources. It is recommended that the report provide the rationale and theoretical basis for the monitoring design chosen to address cumulative effects detection, including its strengths and weaknesses and explanation for the chosen sampling intensities. It may be efficient to establish a stronger connection with the EIAs done in the area.

The Five Year Report refers to the use of a regional hydrologic water-balance model to estimate changes to stream discharge from developments within a watershed (p. 3-68 and 3-122). On p. 3-115, a recommendation is given for calibrating a regional hydrologic model that has been used in environmental impact assessment in the area. While this could be a useful component within the range of approaches, the reviewers caution about using calibration data that are not free of impact. This highlights the usefulness of control basins. It is recommended that the report provide greater detail about this model—how it was built and how it will be applied. This may be a critical component to the cumulative effects analysis if the larger, more complex scales are to be adequately addressed.

3.1.3 Parameters Chosen for Analysis

The detailed Objective 2a limits the investigation of trends and cumulative effects to two climate and three hydrometric parameters. While this may be necessary in light of limitations in the long-term data, it is probably not satisfactory from the perspective of potential impacts. For example, the omission of groundwater from the monitored variables has been discussed under Objective 1 and is a major concern with respect to the cumulative effects analysis. The modest attention given the Athabasca River system is another source of concern, particularly in light of the lack of long-term data sets for the reaches downstream of oil sands developments. Without a detailed discussion in the report providing what is known about the mechanisms for impacts from oil sands development on aquatic systems and ideally identifying specific hypotheses about cumulative effects given the site specifics of the Fort McMurray area, it is difficult to provide more specific comments. It is recommended that the Five Year Report include a section in Chapter 2 describing how oil sands activities can affect aquatic systems and, in particular, regional hydrology, specifically identifying the mechanisms for impact. This discussion would provide a stronger theoretical basis for the subsequent choice of monitored parameters and hence would assist in monitoring the complete range of relevant hydrologic parameters. It is particularly important with respect to Objective 2 due to the complexity of detecting cumulative effects but would also be helpful in addressing Objective 3. It may also be of assistance to less technical readers.

Given the extent of oil sands development in this region and their intensive use of water resources that all, ultimately, derive from the Athabasca River system, it may be necessary to monitor the Athabasca River in more detail. The cumulative impact on low flows from the aggregate of regional development could have significant impacts on fish habitat. Also, are the stations on the Athabasca River and their data sets sufficient to characterize the changes that may occur given the size of this system and the potential for cumulative effects? It is recommended that the potential for cumulative impacts on the Athabasca River system be discussed along with the strategy in place to identify the effects and link them to cause.

3.1.4 Extent of Appropriate Data

There is concern that the monitoring design seems to have grown in an ad hoc manner, driven by regulatory requirements, and without the benefit of an overarching monitoring

design and analytical plan. The majority of recently installed monitoring stations is located in a small part of the study area. For example, the majority of the recently installed hydrometric and all of the climate stations are located in the Tar/Calumet River and Muskeg River watersheds. It is not clear without a monitoring design and analytical plan whether the monitoring station network is sufficient and able to detect an effect.

There is also concern that the hydrometric data record may be inadequate for detecting an effect. According to Table 3-10 (p. 3-28), the long-term data set consists of one climate record of 58-years' duration, one streamflow record of 44-years' duration, and six streamflow records of between 26- and 29-years' duration. All but one of the data sets have less than 16 years of continuous annual data because between 1987 and 2001 all stations but one do not have data from November through February (representing 33% of each year). In addition, there is only one long-term station on the Athabasca River and it is located upstream of most of the developments, just north of Fort McMurray. Given the variability of hydrologic data in general, and these data in particular, these data records may be inadequate for detecting effects, especially cumulative effects, over the range of space and time scales that must be considered by this objective.

There is a specific concern with respect to the record of air temperature. Data from the Fort McMurray station indicate a shift as of 1971 and yet this is prior to the start of six of the seven long-term stations. It is recommended that the 1971 temperature shift be fully assessed so that trends can be adequately understood. There may be a simple reason for the shift that is held in Environment Canada or other published data. Regardless, how does this shift affect the trends and the analyses themselves?

3.2 Methods

The methods used to sample and record climate and hydrometric data appear appropriate. Although there are some data gaps, this is typical and understandable given the harsh northern Alberta environment. The statistical methods used to analyze the data are discussed in the following section. Methodological issues concerning the monitoring design were discussed in the previous section.

3.3 Statistics

Statistical analysis for trends and cumulative effects consists of the repeated application of the Spearman Test for Trend to eight long-term data sets from Environment Canada. The reviewers have a number of concerns with the approach and how it was carried out and reported.

3.3.1 Statistical Considerations in Conducting Trend Analysis

The Spearman Test for Trend is applied repeatedly to the long-term data sets to identify the presence of trends in the climate and hydrologic data over the past decades. In some cases the data are discovered to possess serial dependence which may violate the trend result. Serial dependence reduces the independence of the data and this in turn reduces the alpha and/or beta levels for the test, reducing the significance of the result. The serial dependence is not addressed in any of the trend analyses, nor are corrections or alternatives to the Spearman Test discussed. One suggestion provided here to correct for

the serial dependence is to partition the annual data record into sub-periods that are treated as distinct populations. These subsets can be tested for differences (e.g. between three populations: early, middle, late). It is recommended that serial dependence be addressed in the trend analysis. It is also recommended that alternatives to the Spearman Test for Trend be discussed and, if appropriate, applied.

Exogenous influences are ones in which an additional variable has an important influence on the trend under study and obscures the trend through time under analysis. Exogenous variables are suggested and identified yet analysis is not carried out to address them: for example, the influence of precipitation and temperature on discharge. For the Beaver River, a temperature trend is observed in the data and used to infer an explanation for the water yield trend at one of the long-term hydrometric stations; however, an analysis is not provided to investigate this hypothesis. Removal of confounding effects may remove statistical noise sufficiently to identify real trend signals present in the data, should they exist. It is recommended that analyses be carried out to address the existence of exogenous variables so that the presence/absence of time trends can be adequately assessed.

3.3.2 Reporting of Power (beta) and Statistical Significance (alpha)

The statistical power of the trend detection is not discussed. The report identifies an absence of trends when it is unclear whether the statistical tests applied to the data can detect the trend if it does exist. Not finding a trend does not mean that a trend is absent. This is what the power of the test indicates yet this information is not provided. Power is a function of alpha level, effect size, sample size, and variance (Peterman, 1990). Given the variability inherent in these hydrologic data, the power may be low and hence warrants a discussion and likely a re-analysis. It is recommended that statistical power of each trend test be determined and presented in the report.

The report acknowledges that there are insufficient data to complete a rigorous statistical analysis of the short-term climate and hydrometric data. It is not stated, however, how much data will be required before a statistical analysis with sufficient power or confidence can be generated. It is suspected that several decades of data will be required at each station before there is reasonable confidence in the data and the statistical power is adequate to reach conclusions concerning trends and effects. It is likely that many of the oil sands development reserves, as currently defined, will be exhausted and the landscape reclaimed before the monitoring data will be of much use in verifying EIA predictions. It is questionable then that these data should be collected at all if they will be unable to detect change. It is recommended that a power analysis of the monitoring design be conducted to determine whether the monitoring network that currently exists will be able to detect an effect (and to what degree) if it is present.

The alpha level (statistical significance) of the results is generally presented as significant (90%) or highly significant (95%). It is recommended that the actual alpha level be provided so that the reader can make a more informed interpretation of the outcome of the statistical tests.

3.3.3 Parametric Analyses

The concerns for the trend analyses above (robustness to serial correlation, use of explanatory covariates to reduce variability, power computations) can all be better addressed using parametric (regression and autoregressive) methods. Although parametric linear regression models only capture the linear component of any trend, they nevertheless would permit more flexible, robust and insightful data analyses than non-parametric analyses for purposes of planning monitoring designs. It is recommended that suitable parametric linear model analyses for trend be applied to the hydrological data. These analyses should be directed at determining the sample sizes needed and effect sizes detectable in an achievable monitoring design.

3.4 Results

The section entitled "Conclusions and recommendations" (3.3.4) is essentially a summary of Section 3.3. It focuses almost entirely on the detailed Objective 2a, namely identifying trends in the five hydrologic and climate parameters. The trends and serial dependence are repeated without a conclusion or interpretation provided about the validity of the results. Also, the summaries would be easier to read if they were tabulated. The major objective of determining cumulative effects is not addressed at all in Section 3.3.4, nor in the other concluding sections associated with this objective (e.g. 8.1.2). Overall, given the concerns raised above, the conclusions presented (p. 3-119 to 3-120) do not logically follow from the analyses presented. It is recommended that the Five Year Report include comment on the implications of the findings for achieving the major and detailed objectives.

4.0 Monitoring To Verify EIA Predictions

The program assessed in this section addresses the third primary objective, Monitoring to verify EIA predictions, for the climate and hydrology section. The report discusses the location of RAMP monitoring stations installed since 1997 to monitor both climate and hydrologic parameters in watersheds in the vicinity of or contained wholly within oil sands development areas. Data retrieved from the climate and hydrometric stations are analyzed for temperature, precipitation, water yield, flood discharge, and low-flow discharge. Recommendations are made in the Five Year Report that will increase both the quantity and quality of data collected.

This primary objective is further defined as "collecting data against which predictions contained in environmental impact assessments (EIA's) can be verified". The more detailed objectives, as established by the technical subcommittee are defined as follows:

- to characterize the behaviour of the smaller local areas (streamflow and precipitation) monitored by RAMP and assess their likely behaviour in the longer term; and
- to evaluate whether EIA predictions can be evaluated at this time and whether the data collected by RAMP will be appropriate to do so in the future.

The specific reframed objectives sharpen the primary objective by providing clearer statements of what will be addressed under the topic of "Monitoring to Verify EIA Predictions". However, the reframed objectives limit the scope of this overall section since monitoring to evaluate EIA predictions is limited to an assessment of whether they

can be evaluated now and in the future. Although the following review comments acknowledge and make use of the information provided in these more specific objectives, the review assumes that the broader objective is also to be met, i.e. monitoring to evaluate EIA predictions.

The reviewers assessed this objective for the adequacy of the monitoring design, the methods used, the statistical analysis and the results and conclusions reached. Comments that fall under each of these subjects are found below. For the statistical issues, the reader should refer to Appendix III.

4.1 Monitoring Design

The monitoring design program is meant to satisfy the detailed objectives described above. The reviewers are concerned about several aspects of the monitoring design of RAMP for the short-term stations. These concerns involve:

- linkage between EIA predictions and the monitoring network;
- rationale for the design of the monitoring network; and
- monitoring of low flows.

4.1.1 Linkage Between EIA Predictions and the Monitoring Network

The areas monitored by RAMP appear to represent smaller local areas reasonably well. In addition, the network of RAMP climate and hydrometric stations that is not associated with any particular oil sands development appears appropriately located in order to characterize pre-disturbance hydrology of smaller local areas. Whether stations associated with particular oil sands developments are established in the correct locations to verify EIA predictions, however, is unclear. The reviewers found it very difficult to assess whether each station is located in the correct place when the EIA predictions that require verification are not included in the report and the monitoring design is not clearly presented. For example, as outlined in the Shell Jackpine Phase 1 EIA, oil sands mining will likely excavate into the Pleistocene Channel Aquifer. Will the RAMP hydrometric Station S2 located on Jackpine Creek be able to monitor and detect the changes to surficial hydrology due to this activity? More generally, what water-related issues outlined in EIAs require verification, how will a station location and sampling rate detect changes to surficial waters as predicted in the EIAs, and how much data over what period are required in order to detect an effect at each station? (See discussion in review of Objective 2 on statistical power analysis.) These questions indicate the level of analysis and discussion that is required in order to satisfy the second detailed objective. Thus, it is recommended that a detailed discussion that identifies the variables that are likely to be impacted and the magnitude of the impact that it is necessary to detect be provided. Based on this discussion, the variability, controls, sample sizes, etc. that enable detection can be determined. It is also recommended that, where applicable, the linkage between monitoring station location and operation and relevant EIA predictions be detailed, discussed and analyzed where possible. Only after these discussions take place will it be possible to determine whether EIA predictions can be verified. The lack of discussion over what specific EIA predictions the monitoring network is attempting to verify and how the current and proposed monitoring network will be able to evaluate EIA predictions is one of the weakest but most important aspects of this section.

4.1.2 Rationale for the Design of the Monitoring Network

In order to assess whether the monitoring program design is adequate to evaluate EIA predictions, it is necessary to understand the monitoring design rationale. Although the monitoring design rationale is presented in Table 2.2 of the RAMP Program Design and Rationale (RAMP, 2002b) and in p. 3-70 to 3-75, detail and discussion are lacking. For a monitoring program the size and importance of RAMP, a lengthy discussion, even a separate report, outlining the monitoring program design rationale is required. Such a report would include details concerning station location rationale, history, location limitations, geomorphological features present in the watershed, watershed response to precipitation, how the station location suits the data requirements of RAMP's components (benthics, sediment, water quality, etc.), how the station complements the regional monitoring objectives and requirements, etc. Without this kind of background information, it is impossible to determine whether the RAMP monitoring stations are located correctly, are sampling at a sufficient rate and what additional stations are required at what locations in order to effectively and efficiently monitor effects and to evaluate EIA predictions. See Objectives 1 and 2 for additional discussion.

4.1.3 Monitoring low flows

Several of the short-term or recently installed hydrometric stations do not sample low flows during the winter season, on the assumption that some of the smaller monitored streams freeze to the bottom. Monitoring low flows, however, is important as the quantity and quality of low-flow discharge often dictates survival of aquatic biota, including fish over-wintering periods. Since oil sands development is likely to affect low flows, it is important to monitor these changes. The environmental impact of having winter flows reduced to zero could be extremely significant for a stream that typically experiences low to very low flows. It is recommended that all reasonable efforts be made to provide continuous sampling and recording of winter flows at all RAMP hydrometric stations, regardless of known or assumed winter discharge characteristics.

4.2 Methods

The methods used to analyze the data appear appropriate. However, there are some issues of concern that should be addressed in future reports. These issues are detailed in the template report found in Appendix IV.

4.3 Results

The results of the analysis are adequately presented, although the accuracy of some of the results is of concern, given the comments above and presented in Appendix IV. There is also a good general description of each watershed’s geomorphology and the hydrologic response that can be expected. The variability on an annual basis of both climate and hydrometric data is also reported satisfactorily. What is absent from the report, however, is a detailed variability analysis on a monthly basis; for example, between all of the June records or all of the July records. There is also need to report figures and statistics such as variability between months, during periods important to fish and other aquatic biota, between peak flood dates, between low-flow events, etc. There are many variability statistics useful in charaterizing watersheds pertinent to biological processes that are not

included in this report. Although the reviewers acknowledge that there are few data with which to work, this additional analysis would help to characterize the behaviour of smaller local areas. It is recommended that the parameters that require variability analysis and comparison (i.e. monthly data, month-to-month data, low-flow periods, etc.) be identified and a thorough variability analysis and comparison be completed and reported.

5.0 Recommendations and Suggested Implementation

5.1 Overall Assessment

The report is written in a manner that is accessible to most stakeholders, with a few minor exceptions. For the most part, the report section is well organized and clearly written. The reviewers acknowledge that it is difficult to write a report to be readable and acceptable to all stakeholders and reviewers of differing backgrounds and understanding of environmental monitoring practices. In order to reach a broader audience, many terms, particularly statistical terms, should be defined and their usefulness in charaterizing natural variability, including strengths, weaknesses and limitations of tests outlined. This information could be included in a glossary in an appendix. Researchers and decision-makers would have a difficult time using this report because it does not provide enough technical detail, depth of analysis and discussion around pivotal issues. It is recommended that the report audience be defined in the introduction, and additional information be included in an appendix for readers outside of the defined audience. It is also recommended that brevity be enhanced through a greater use of tables to avoid repetitive text where possible.

With respect to the three objectives, the reviewers have a number of concerns with the hydrology and climate section. For the first objective, most of these concerns deal with the lack of information on the analytical plan and monitoring design in addition to statistical analysis, statistical error associated with data interpolation, and the limited scope of the variability analysis. The background information provided for each watershed is informative and the explanation of how and why different watersheds react to precipitation inputs differently is helpful.

In assessing the second objective, the reviewers recognize that trend and cumulative effects analyses are demanding in long-term data requirements and that new programs are limited, to some extent, by what has been done before. RAMP has looked at the data sets available and begun an analysis relevant to this objective. A number of shortcomings were encountered in reviewing the second objective. A coherent analytical basis for the cumulative effects analysis was not provided in the Five Year Report. In addition, it is very difficult to evaluate the monitoring network distribution until the monitoring design is clearly presented. Some definitions are ambiguous or not provided, leaving the reader unclear about what is being monitored. Some statistical tests are incomplete. As a result, some of the conclusions reached are inappropriate at this time. Additional key background information relevant to the objective, greater attention to detail, and a presentation of the strengths and weaknesses of the data sets (in relation to the objective) would strengthen the section.

Overall, the section on the third objective was written well but there are several key issues that are either omitted or need additional discussion and review. Specific EIA predictions are not presented, evaluation of the monitoring program is incomplete, and data variability analysis is inadequate. The report makes many recommendations to increase temporal and areal data collection abilities. The reviewers concur with many of the recommendations concerning monitoring station upgrading that are mentioned in this section, especially to measure low flows during winter.

5.2 Gaps

In light of the concerns outlined above, the reviewers have identified several major gaps in the climate and hydrology section:

- lack of a detailed monitoring design and analytical plan;
- significant data gaps over the low-flow period, the time most critical for many aquatic biota;
- lack of a detailed monitoring design and analytical plan;
- significant data gaps over the low-flow period, the time most critical for many aquatic biota;
- limited data variability and comparison analysis;
- absence of statistical power reporting in trend analysis;
- absence of a strong analytical framework for monitoring and detecting cumulative effects;
- lack of long-term data sets including absence of certain parameters;
- lack of linkage between monitoring station location and relevant EIA predictions that are meant to be verified by monitoring stations;
- need for monitoring of low flows at all hydrometric monitoring stations, regardless of known or assumed discharge characteristics; and
- incomplete consideration for the cumulative impacts to the Athabasca River system.

The reviewers recognize that the basic objectives are ambitious and difficult to meet and, as a result, gaps and program weaknesses are to be expected. RAMP has begun the job of assembling information sources for addressing hydrologic impact. Some of the above gaps may be dealt with in subsequent annual reports by including new material to expand on what has already been presented. In other cases, the gaps point to new areas that the RAMP will need to move into if the basic objectives are to be met. Many specific recommendations are provided below.

5.3 Recommendations

Recommendations have been provided throughout this review pertaining to the three primary objectives. Additional recommendations and discussion of the recommendations below are found in Appendix IV. The following is a summary of the most significant recommendations for each objective. For each recommendation given, the section number is provided where the detailed rationale can be found (in this report).

Objective 1

Five primary recommendations are outlined below in order of priority.

1. Provide a detailed discussion in subsequent monitoring reports outlining the monitoring design principles, objectives, and how the current and proposed monitoring network addresses those objectives (2.1.1).
2. Identify time periods and parameters of interest for variability analysis and complete a thorough variability analysis and report (2.4).
3. Monitor flows from November through February for all hydrometric stations, regardless of known or assumed historical or current discharge characteristics (2.1.3).
4. Report the statistical confidence level and illustrate using confidence bands on all figures indicating extreme events (2.3.1).
5. Include local and regional groundwater monitoring by data acquisition from oil sands developers; place groundwater monitoring stations in areas considered environmentally sensitive and outside of oil sands developers' existing and proposed monitoring areas (2.1.2).

Objective 2

Seven primary recommendations are outlined below in order of priority.

1. Develop the analytical plan for the cumulative effects analysis and clearly present it (3.1.1).
2. Provide in the report the rationale and theoretical basis for the monitoring design chosen to address cumulative effects detection, including its strengths and weaknesses and explanation for the chosen sampling intensities (3.1.2).
3. Conduct a power analysis of the monitoring design to determine whether the monitoring network that currently exists will be able to detect an effect (and to what degree) if it is present (3.3.2).
4. Apply a suitable parametric linear model analysis for trend to the hydrological data.
5. Discuss the potential for cumulative impacts on the Athabasca River system along with the strategy in place to identify the effects and link them to cause (3.1.3).
6. Include a section in Chapter 2 of the Five Year Report describing how oil sands activities can affect aquatic systems and, in particular, regional hydrology and specifically identifying the mechanisms for impact (3.1.3).
7. Establish control watersheds to act as benchmark comparisons as oil sands developments are further implemented (3.1.2).
8. Determine and present in the report statistical power of key tests (3.3.2).

Objective 3

Four primary recommendations are outlined below in order of priority.

1. Provide an in-depth discussion that identifies the variables that are likely to be impacted and the magnitude of the impact that it is necessary to detect. Based on this discussion, determine the variability, controls, sample sizes, etc. that enable detection (4.1.1).
2. Analyze and describe the linkage between monitoring station location and relevant EIA predictions, as applicable (4.1.1).

3. Identify parameters that require variability analysis (i.e. monthly data, month-to-month data, low-flow periods, etc.) and conduct a thorough variability analysis and report (4.3).
4. Provide continuous sampling and recording of winter flows at all RAMP hydrometric stations, regardless of known or assumed flow characteristics (4.1.3).

General Recommendations

In addition to these specific recommendations, the following are also provided:

1. Include in the report a glossary providing a clear expression of the meaning of technical terms (3.1.1).
2. Provide greater detail about the regional hydrologic water-balance computer model used for EIA predictions—how it was built and how it will be applied (3.1.2)?
3. Define the report audience in the introduction and include additional information in an appendix for readers outside of the defined audience. Enhance brevity through the greater use of tables to avoid repetitive text where possible (5.1).

5.4 RAMP 2002-2009 Plan

The reviewers have included many recommendations throughout the review of the climate and hydrology sections. These recommendations should be considered in the development of future RAMP monitoring activities. However, as stated throughout this review, the need for a monitoring design and analytical plan is apparent. No additions or modifications to the RAMP climate and hydrology monitoring program should be made without first developing a monitoring design and analytical plan.

5.5 Concluding Remarks

The review summarized in this report highlights the major gaps in RAMP preventing the objectives from being met. By carrying out the recommendations provided, the key gaps can be addressed. Where it is not possible to address certain recommendations, it may be preferable to adjust the RAMP objectives to reflect what is possible given the data sets involved.

ASSESSMENT OF WATER QUALITY COMPONENT
PREPARED BY NANCY GLOZIER, JAN BARICA AND MONIQUE DUBÉ

1.0 Introduction

The Regional Aquatic Monitoring Program (RAMP) "…was designed as a long-term monitoring program that incorporated both traditional and scientific knowledge" (p. 1-2, RAMP, 2003). RAMP is a multistakeholder program composed of funding (oil sands industries) and non-funding (regulators, First Nations, NGOs, and local communities) participants with membership having evolved through the five-year period since 1997. Its mandate is substantial; specifically, "to monitor, evaluate, compare, review and communicate the state of the aquatic environment in the Athabasca Oil Sands Region" (p. 1-4; RAMP, 2003). In addition to documenting changes in aquatic ecosystems over time, an objective within RAMP was to determine if observed changes were caused by natural variability, cumulative effects of development, or both. With the Oil Sands Region experiencing rapid growth from 1997 to 2001, annual modifications were made to the monitoring program. These changes affected RAMP's organizational structure, objectives, study area, and study design. This chapter deals with the water quality monitoring program. A major issue that arose in the review of the water quality section was that the frequent changes associated with the program over time and space made it very difficult to get a sense of what was measured, where, and when.

The three reviewers of the water quality component have extensive experience in study design, and analysis of water quality data. The review concentrated on Chapter 4 of the Five Year Report, with additional reference to Chapter 8 (Conclusion and Recommendations), the annual RAMP reports (1997-2001), the Oil Sands RAMP Program Design and Rationale, and the Biostatistics Review of RAMP (Appendix III). These documents were reviewed in the context of the RAMP main and sub-objectives of the water quality program. The main objectives included: characterizing existing variability, detecting and assessing cumulative effects and monitoring to verify EIA predictions. The sub-objectives included: influence of river discharge on water quality, fall vs. winter sampling, spatial trends in the Athabasca River, correlation between parameters, and duration of sampling for establishing baseline conditions. Finally, the 2002-2009 Program Design Document was reviewed to determine if the proposed design identified gaps and issues, subsequently improving the RAMP program to ensure that the main objectives were being addressed. This chapter is a summary of the water quality template, which appears Appendix IV, and to which the reader is directed for greater detail.

2.0 Characterizing Existing Variability

The water quality (WQ) component attempts to "characterize existing variability" through three sub-objectives: (1) parameter correlations, (2) examination of data relative to changes in discharge and (3) comparisons of parameter concentrations observed in fall vs. winter sampling. There appears to be confusion throughout the report and across the major sections (water, fish, benthos) on why and how to meet the objective of characterizing variability. The intention should be to develop an understanding of the range and magnitude of key indicators of water quality, specifically in relation to

potential effects from oil sands development. Documenting the existing natural variability allows for future comparisons of the effects of developments (i.e. we would have the "baseline" to evaluate the importance of any shift in indicator values resulting from landuse changes). Unfortunately, the RAMP WQ program has varied so significantly over time it is difficult to determine if data currently exist to characterize variability for any of the aquatic systems potentially impacted by oil sands development.

Specifically, the parameter correlations performed in the report show little relevance to characterization of variability that could be used in the future as baseline information. The data are available, but are not presented in a manner consistent with documenting baselines. Although examining general correlations among parameters from the RAMP parameter list is a general/universal approach for any water quality monitoring program, rationale on parameter importance to the RAMP program is lacking. For example, principal components analysis (PCA) was done on conventional parameters (nutrients, major ions, and 19 metals (16 of them as dissolved), in no order of significance of their potential impacts. The parameter correlations should have ultimately identified key indicator water quality parameters that can be used to monitor change due to oil sands activity. A desktop exercise was required as a first step to list which parameters are currently being monitored, which are regulated and for what purpose, which are used in the environmental impact assessment (EIA) predictive models, which parameters are expected to change with development, and which parameters currently have site specific objectives or Canada Council of Ministers of the Environment (CCME) criteria for the protection of freshwater aquatic life. From this desktop inventory, a more focused analysis could have been conducted. Nowhere in this section is there a recommendation on which parameters should be measured in the future because these parameters are the most suitable indicators to characterize ecosystem variability. Additional information that would contribute to characterizing variability would include spatial and seasonal patterns within and between appropriate groups of reference sites.

Two further exercises performed under the objective of characterizing variability were (1) determining the relationships between river discharge and (2) parameter values/variability in fall and winter. Although these relationships are interesting, they present nothing unexpected and have no direct relevance to oil sands environmental concerns. Correlations of parameters to the river flow rates (Figures 4.8-4.9, Table 4.14) are expected; as in any river, major ion concentrations in the Athabasca River increase during the periods of low flow (winter). These relationships do provide useful hydro-geochemical information, and should be published separately in a science journal. However, the main objective of these two exercises is unclear. It is assumed that the authors want to determine when the period of maximum impact might be (or the period of greatest sensitivity) and monitor accordingly. Consideration of the appropriate time to sample depends upon the activities of the development, logistics, and when the other aquatic components are being measured. What if significant differences in water quality are observed downstream of development in the winter? The next question will be: Are those changes affecting biota? This highlights a key factor missing from the RAMP program, i.e. the linkages between components and the identification of which components are considered "effect" components and which are considered "supporting"

components. Following the national Environmental Effects Monitoring Program (Environment Canada 2001), water quality assessments are considered supporting information for examining effects on biota. This approach should be seriously examined for the RAMP program and, presumably, would lead to a clearer understanding of the role of chemical water quality within the program.

Thus, this objective has not been adequately addressed over the first five years of RAMP.

3.0 Detecting Regional Trends and Cumulative Effects

The second objective, detecting regional trends and cumulative effects, was addressed in the RAMP WQ section using three approaches: (1) temporal trend analysis, (2) examining spatial patterns/trends between sites, and (3) with power analyses to determine the ability to detect changes in water quality.

Temporal trends were examined for water quality data collected on the main stem of the Athabasca and Muskeg Rivers. Alberta Environment's database on the Athabasca River allowed for long term (1976-2001) temporal trend analyses at two sites: (1) upstream of Fort McMurray and (2) far downstream at Old Fort. Shorter term (1997-2001) temporal trends were examined for two Muskeg River sites (upstream and downstream). Seasonal Kendall tests and Sen's slope were used for trend analysis (WQ StatPlus). Although some trends were detected, discussion regarding the importance of the variables that changed with time or their relevance to expected changes with oil sands development is lacking. Additionally, direct or indirect comparisons of the magnitude of changes through time and between sites are not discussed. Finally, there is no basis given for selection of the two aquatic systems assessed or discussion of how they factor into the existing and proposed development on these systems over time and space.

Spatial trends were examined for the same sites, upstream and downstream on the Athabasca River and Muskeg Rivers, as well as additional sites on tributaries and in wetlands. Overall spatial patterns between these sites were determined using PCA. The differences between the main stem of the Athabasca River, its tributaries, the Muskeg River, and wetland habitats are interesting but not particularly surprising. The point of spatial analyses within RAMP should be to determine if differences relative to locations of oil sands development exist, not to compare different aquatic ecosystems to each other. Some relevant points missed in these analyses include the establishment of baseline conditions and which aquatic ecosystems have similar chemical/physical characteristics. These similarities could then be linked to similarities/dissimilarities in the biotic community assessments.

A major conclusion in this section, stemming from comparisons of the two Alberta Environment sites on the Athabasca River (separated by >150 km), that "cumulative development in the oil sands area had not resulted in the degradation of water quality within this stretch of the river" (p. 4-52, Section 4.3.1.3) is not warranted. The single downstream site on the Athabasca River is ~90 km downstream of current oil sands activity and there are many confounding factors, apart from any changes due to the natural river continuum, to warrant this conclusion.

The work on the Muskeg River is the first indication that there was a sampling design suitable to measure changes due to oil sands development. However, the direction this section takes is confusing; observed differences in sulphate are attributed to discharges from the Alsands Drain but then it is stated that cause-effect is unknown. The authors do not assimilate this information or establish it as a baseline for future assessments. The next questions could have been: What is the magnitude of the change (i.e. how far downstream does it go) and what are the biotic community response patterns in this aquatic system?

An assessment of the ability to detect a certain magnitude of change in water quality parameter values was discussed. In regard to the nonparametric trend analysis, the recommendation of four sampling years was based on the software recommendations for these analyses. This is certainly a consideration; however, a recent publication on the design of water quality monitoring programs to detect trends (Vecchia, 2003) should be consulted for confirmation. The number and seasonal placement of samples depends on the pattern and variability of water quality characteristics within the watershed.

For ANOVA-type analyses a series of power analyses were conducted to determine the magnitude of change in a parameter the current program would be capable of detecting (i.e. the effect size). For some parameters (e.g. total boron) a large change (228%) would be undetectable, whereas for others (total dissolved solids, TDS) a small change (6%) between sites would be statistically detectable. This is important information and should influence redesign of the monitoring program. However, the first step is to establish the core parameters of concern influenced by oil sands development and an acceptable level of change linked to effects on biota. Once this desktop exercise is complete, then power analyses can be most useful in redesigning the RAMP WQ program.

Unfortunately the current monitoring design for this objective is not adequate to measure cumulative change related to oil sands development. However, with the background information now available, an excellent opportunity exists to improve RAMP with clear objectives established.

4.0 Monitoring to Verify EIA Predictions

It was assumed that this section would summarize the current EIA predictions, review the available information from RAMP, and compare these predictions and results to determine the accuracy of the EIA predictions. However, the fundamentals of the EIA were not summarized or even cited anywhere within the Five Year WQ Report. Instead, this section attempted to answer the following questions:

1. Are the sampling locations appropriate to evaluate the EIA predictions when development actually happens?
2. Are the water quality parameters appropriate to evaluate the EIA predictions?
3. Is the appropriate type of information being collected to detect human influences?

Therefore it is difficult to assess whether this section meets its objectives of verifying the EIA predictions. The following comments pertain directly to how the RAMP WQ report addressed the three questions posed.

Discussion regarding appropriate sampling locations revolved around presenting arguments as to why the sampling locations are valid for EIA. An alternative approach would have included a discussion on the limitations in the current sampling locations and suggestions of priority areas that need further examination for EIA evaluations. For example, the EIA nodes on the Athabasca River are located downstream of the tributaries. However, the RAMP WQ program monitored the Athabasca River upstream of the tributaries. The authors conclude that the latter approach was sufficient because, even though they were trying to examine the cumulative impact of a tributary on the Athabasca, monitoring upstream of the tributary confluence "can still be used to monitor potential effects from upstream". However, this would likely not be valid, particularly if there are a multitude of other influences in between. The standard approach would be to monitor both upstream and downstream of the tributary confluence with the Athabasca River.

Additionally, the validity of the statement that "inclusion of the upstream station of the Embarras River site near Old Fort permits potential verification of cumulative development within the basin (p. 4-72)" depends entirely upon your definition of cumulative. The goal of an EIA is to monitor the cumulative impacts of oil sands development. That means examining the effects of developments in isolation and in combination to determine if changes are localized or if they begin to accumulate in additive, synergistic, etc. fashion. This requires a systematic, spatially and temporally iterative approach to monitoring. Monitoring one site 165 km away may, over the long term, show changes but there will be no mechanism to determine if those changes were due to development, climate change, or just the normal changes a river goes through over time and as part of the natural river continuum. We completely disagree with the author's assessment of the program's ability to measure change.

Parameter lists are apparently complete; however, more discussion regarding consistency across all sampling programs would improve the analyses. Regarding nutrient analyses, no particular forms were measured nor were totals. These would add additional comparisons outside of RAMP. The parameter list needs a complete focus to a consistent core, consistent with Alberta Environment and focused on what is essential to understand the fundamentals of WQ and what indicators you would expect to change with oil sands development. Finally, a comment must be made regarding Table 4.1. Much of our time and effort was absorbed in attempting to clearly understand the water quality monitoring program. Although it is recognized that there are complexities, numerous changes, and numerous agencies involved, Table 4.1 did not elucidate the strengths/weaknesses of the program. The vast array of symbols (15) used to indicate which parameters or combination of parameters were sampled, at which site/time, disallows use of the table in an easy and transparent manner and precludes the review of the table for one particular parameter type. For example if one was looking for all the sites and times for which

PAHs were sampled, there are 6 independent symbols for which PAHs were included as part of a unique combination.

The RAMP program provides a significant opportunity to illustrate how baseline and follow-up monitoring can be done in a consistent way over time. The ultimate objective of determining if the EIA predictions were accurate, adaptively managing the system if they were not, and developing a process and database to improve predictive models and monitoring in the future are all real possibilities of this program that have yet to be realized.

5.0 Recommendations and Suggested Implementation

The RAMP Five Year WQ Report is an exhaustive document that contains a large amount of valuable information. The RAMP writing team did a reasonable job attempting to compile such an enormous amount of information. However, some severe editing, eliminating frequent repetitions, and condensing the report size would greatly improve its quality and accessibility to the key messages.

RAMP and its stakeholders should be commended for their willingness to participate in an external review of the program as well as for their willingness to work together over a regional scale. Clarification of the mandate and objectives of the RAMP WQ program, however, is required before further interpretation is completed. The lack of a clear purpose/roll for the WQ monitoring program within RAMP likely contributed substantially to the majority of the issues raised in this review. It should be recognized that the three primary objectives of the RAMP WQ program (and RAMP overall) are interdependent. The overall goal is to synthesize on an on-going basis what the original EIA impact predictions were and, through a well-designed monitoring program, determine if those impact predictions were accurate. Getting to this stage requires characterization of variability as well as on-going measurement of spatial and temporal trends and cumulative effects.

Overall, RAMP has enormous potential to serve as a national and international example of integrated, multi-stakeholder monitoring. Unfortunately the WQ component of the program falls severely short of the three main RAMP Objectives, based upon the annual reports and the Five Year Report. It is clear that large volumes of data exist and certainly analyses of these data could be repeated with clearer questions and greater focus. That being said, after five years and considering the development pushing ahead in the oil sands, it is alarming that the main monitoring program for the area significantly lacks strategic direction and scientific process. In the current state and based on the annual reports and the Five Year Report, the RAMP WQ program is not in a position to measure and assess development-related change locally or in a cumulative way.

The major gaps of this component are as follows:

1. There is not a strategic process for establishing sampling locations or for addressing the three primary objectives in an organized, focused and science-directed way.

2. There is no integration between WQ and other RAMP components and a lack of understanding of the role of WQ in RAMP. Is the WQ program a supportive component to the biotic component or an effect endpoint in and of itself? The former would be consistent with other Canadian monitoring programs.
3. There is a lack of core consistency for parameters measured, analyses conducted, statistics conducted, and reporting of results.
4. There is a lack (or insufficient knowledge) of specific markers or WQ indicators of oil sands development.
5. The study design does not build upon well-established, state-of-the science knowledge in Canada and elsewhere.
6. The current method of result dissemination and reporting is not sustainable. An information management and assessment system is required that builds off similar initiatives in the region.
7. Although there has been cooperation with provincial monitoring programs and other scientific programs such as PERD and perhaps NREI, these reports are not reviewed or provided in the Five Year Report.

Major recommendations for improving the WQ program within RAMP are divided into two components: (1) study design and (2) integration/management.

1. **Study Design.** A strategic overhaul of the RAMP WQ monitoring program is required in conjunction with a review of the other RAMP components. Revisions should include development of a strategic sampling plan, selection of a core parameter list including detection limits and analysis methods and core reporting requirements. The sampling plan and selection of core parameters should be directly related to the location and nature of existing and proposed developments. The parameters selected for the current RAMP are not oil-sands-development specific, but of a generic type, used by most WQ monitoring programs. Selection of key parameters should be done in view of the RAMP results to date and "markers" of oil sands impacts highlighted and expanded. Consideration of winter sampling for specific reasons (e.g. in areas of development) could be considered, but should not be at the loss of the core autumn sampling. The program should also build upon existing success stories that are established and proven outside of RAMP (e.g, EEM, effects-based monitoring). Finally, it is imperative that the WQ program not be conducted in isolation to the other RAMP components (benthos, fish) but rather as an integral part of an integrated site assessment. The current program stretches too far and wide at the expense of replication and consistency. The panel design proposed by B. Schwarz (Appendix III) should be considered as well as the recent US Geological Survey (USGS) document regarding sampling design for WQ monitoring programs (Vecchia, 2003). A clear opportunity exists for RAMP to utilize the large volumes of data available to create a world-class, science-based cumulative effects monitoring program.

2. **Integration/Management**. The component-based approach to RAMP (water group, benthic invertebrate group, fish group) has led to fragmentation and a lack

of integration. Consideration should be given to dissolving this management structure, or at the very least, developing an integration team that also serves as a scientific advisory panel. RAMP has been severely limited by the many of changes in the program.

An information management and assessment system should be considered because the existing assessment and reporting process is not sustainable. This system should consider and build from other initiatives in the area and consider inclusion of provincial, federal, industry (e.g. oil and grease measurements), and RAMP data. This information system would provide key plots and analyses on a consistent basis over time for all components. Location of sample stations on a GIS-based map and relative to existing and future development is also required. It is too difficult to track where water, benthic invertebrate, sediment and fish samples were taken because the program has changed so frequently.

Finally, we gave an overall ranking of unsatisfactory for the WQ program. We wish to make it clear that this ranking does not pertain to the actual Five Year Report itself but to the overall RAMP WQ program, how it addresses the three primary objectives, and its current implementation relative to other scientific practice in WQ monitoring in Canada.

ASSESSMENT OF SEDIMENT QUALITY COMPONENT
PREPARED BY BRIAN BROWNLEE AND UWE BORGMANN

1.0 Introduction

RAMP (Oil Sands Regional Aquatic Monitoring Program) began in 1997 as an aquatic monitoring program within the area of oil sands development in northeastern Alberta. We have reviewed the sediment quality sections of the Five Year Report covering the period 1997-2001, and portions of the sediment quality sections of the annual reports from 1997 to 2002. More detailed review comments are Appendix, IV.

Three RAMP objectives were evaluated:

1. Characterizing existing variability
2. Detecting and assessing cumulative effects and regional trends; and
3. Monitoring to verify (test) environmental impact assessment (EIA) predictions.

We reviewed both the Five Year Report and the program. In the case of sediment quality, for some objectives, the Five Year Report did not do justice to the program. Accordingly, we referred to the annual reports to gain a better understanding of what the program was doing and accomplishing. For the first objective, we found it helpful to distinguish between the program and Five Year Report in the template reviews. Relatively minor changes are recommended for the program, but major improvements are needed in the areas of data analysis and reporting for future summaries.

2.0 Characterizing Existing Variability

Recommendations for the program are limited to quality control, within-site variability and expansion of sediment toxicity testing to include bioaccumulation of metals. Until variability is better characterized, there is no reason to increase sampling intensity. Spatial coverage is already extensive, with 36 sites being sampled in 2002.

For some sites and substances, year-to-year variability has been high; for example, total recoverable hydrocarbons upstream from Donald Creek, east bank. Within site (same sampling occasion) and year-to-year variability for a site need to be separately defined and characterized. The closest example we are aware of that may be applicable was done during the Northern River Basins Study. Crosley (1996) collected 10 replicate samples at a number of sites on the Athabasca River. These 10 replicates were separated into coarse and fine fractions and then analyzed for resin acids. Crosley's results may have limited applicability because the nearest site was well upstream of Fort McMurray and samples were separated into fine and coarse fractions.

Toxicity testing was done on nearly half of the sediment samples collected from 1997-2001. We recommend that future work include bioaccumulation measurements for metals, because body concentrations are a better indicator of bioavailability and the cause of toxicity (Borgmann et al., 2001; Borgmann 2003a, b). In the template, we suggested "metals with concentrations close to ISQGs" as a category for data analysis. However, this may be a bit simplistic since ISQGs (Interim Sediment Quality Guidelines) are based on correlations and not on cause-effect relationships. Since the focus should be on metals

in both water and sediment that are most likely to cause toxicity, comparison of metal concentrations in water with water quality guidelines is better than comparison of metal concentrations in sediment with ISQGS. In this area, the water quality and sediment quality components should coordinate.

The data analysis in the Five Year Report tested for substances that co-occur, examined the effect of sediment composition on PAH levels, and looked for indicator "parameters" that would enable reduction in the number of PAHs analyzed. Existing variability was not characterized. Other notable omissions were the lack of use of river hydraulics and sediment transport in discussing the results, and sources such as natural erosion of oil sands were not considered.

3.0 Detecting and Assessing Regional Trends

We question whether Principal Components Analysis is the best way to look at temporal and spatial trends in the region, and the ability to detect change. The Five Year Report was devoted exclusively to Principal Components. We question the value and validity of PCA for monitoring temporal trends. Further discussion can be found in Appendix VI.

Very little mention was made in the Five Year Report about cumulative effects. Sediment toxicity results were not presented or discussed in the Report.

4.0 Monitoring to Verify EIA Predictions

The Five Year Report considered three questions: are the samplings sites in appropriate locations, does the analytical list include all relevant substances and parameters discussed in EIAs, and is RAMP collecting or obtaining the necessary information to distinguish between natural variability and changes associated with human (industrial) activity?

We suggest that a more effective and meaningful way to evaluate RAMP against this objective would be to take recent EIA as a case study. For example:

1. What RAMP data were used in preparing the EIA?
2. How many years of baseline data were available?
3. How will the monitoring programs of RAMP and the project(s) be coordinated?
4. What predictions were made in the EIA?
5. How will the current RAMP go about testing these predictions?
6. Will RAMP be able to detect project-specific impacts?
7. Can cumulative effects in the region be identified?
8. Can RAMP distinguish between natural variability and industrial inputs?
9. Can RAMP identify the effect of sources other than industrial: natural erosion of oil sands, municipal sources, upstream sources, forest fires, etc.?

This may give a good indication of the likely future performance of RAMP in attaining this objective.

5.0 Recommendations and Suggested Implementation

Objective 1 – Program

- Quality Control. Analogous to water sampling, use clean sand or a low-level sediment reference material for field and trip blanks.
- Quality Control. When high concentrations appear at some sites, as in 2000 for total recoverable hydrocarbons (TRH) and many of the PAHs, there should be a procedure in place to double check and confirm that this did not occur because of field or laboratory contamination.
- Within-Site Variability. For statistical purposes, it is desirable to define within-site variability. One possibility is to take a sufficiently large number of samples at one site to define variability.
- Incorporate bioaccumulation measurements for metals in the sediment toxicity testing.

Objective 1 – Five Year Report and Data Analysis

- Redo the data analysis to demonstrate the range of variability for different substances.

Objective 2

- For reasons of temporal trend analysis, the baseline sampling period for new projects should be extended from three to five or more years, as recommended in the Five Year Report. This will require earlier notification of intent by proponents.
- Using the same approach as in the Five Year Report, temporal and spatial trends and the ability to detect change should be analyzed using examples of individual substances or logical groups of substances. The use of Principal Components in the report did not reveal much about the character of the underlying results.

Objective 3

- The most effective way of determining how well RAMP results will support testing of EIA predictions may be to use a recent EIA as a test case.

ASSESSMENT OF BENTHIC INVERTEBRATES COMPONENT
PREPARED BY DAVID ROSENBERG, MONIQUE DUBÉ, AND STEPHANIE SYLVESTRE

1.0 Introduction

The Oils Sands Regional Aquatic Monitoring Program (RAMP) "…was designed as a long-term monitoring program that incorporated both traditional and scientific knowledge" (p. 1-2/3). Its intent was to document change in aquatic communities over time and determine if change was caused by natural variability, cumulative effects of development, or both (p. 1-1/4). RAMP is a multistakeholder initiative composed of funding (oil sands industries) and nonfunding (regulators, First Nations, NGOs, and local communities) members (p. 1-3/2). The Oil Sands Region experienced rapid growth from 1997-2001, the period of review for RAMP, so changes were made to the program annually (p. 1-3/3). These changes affected RAMP's organizational structure, objectives, the study area, and the study design, as will be evident below. RAMP included several subject areas. This chapter deals with benthic invertebrates, the animals that live on the bottoms of lakes and rivers. These organisms are routinely used in biomonitoring the water quality of lakes and rivers (e.g. Rosenberg and Resh, 1993).

The reviewers of the benthic invertebrate component all have extensive experience in biomonitoring using benthic invertebrates. The review concentrated on Chapter 6 of the five-year review document, with additional reference to the annual benthic invertebrate reports produced during the 1997-2001 review period. The review mainly focussed on the three objectives enunciated in the Five Year Report (p. 6-6 to 6-7): (1) "collecting scientifically defensible baseline and historical data to characterize variability in the oil sands area"; (2) "monitoring aquatic environments in the oil sands area to detect and assess cumulative effects and regional trends"; and (3) "collecting data against which predictions contained in environmental impact assessments (EIAs) can be verified". This chapter is a summary of the more detailed benthic invertebrate template, which appears in Appendix IV at the end of this report. **The chapter and the template are meant to be read together** because there are several cross-cutting issues not specifically identified in sections 2-4 below, which appear in Section 5 (recommendations).

2.0 Characterizing Existing Variability

This objective was broken into two subobjectives: (1) spatial variation in benthic community structure, and (2) baseline ranges for key benthic community variables. The first subobjective was an exploratory analysis of patterns in benthic data from historical and 1997-2001 sampling, and an attempt to identify environmental variables driving those patterns. The results were largely inconclusive, and no specific recommendations were made. The second subobjective was an attempt to characterize variability by establishing baselines for a number of invertebrate metrics. This part was only marginally successful because of the disparate data involved and the short sampling period. The development of critical effect sizes to be used in future evaluations of monitoring data was recommended.

Not much directly pertinent to the detection of development-related change was delivered in the examination of this objective. The fault lies with the naïve nature of the objective. Biomonitoring approaches are currently being used that incorporate variability as part of the way they are done; there is no need for separate studies of variability. Moreover, given the disparate database, it is not surprising that the analysis was largely futile.

The objective should be reoriented around detecting development-related change. Use of already existing biomonitoring programs such as Environmental Effects Monitoring (EEM; Environment Canada, 1997, 1998, 2001; Glozier et al. 2002; Walker et al. 2002; Dubé, 2003) or the Reference Condition Approach (RCA; Reynoldson et al., 1995; Rosenberg et al. 1999; Wright et al. 2000; Bailey et al. 2004); would do this. Use of EEM or RCA would also solve other major problems: (1) standardized data collection (i.e. sampling the same sites over time, using consistent sampling gear and mesh size), (2) use of critical effect sizes and core effect endpoints (these items have to be designed and included at the outset of the program), and (3) provide reference site/area data (sadly missing from the present work).

Examination of the objective could have been improved had the extant EEM program upstream on the Athabasca River been accessed.

Our difficulty in trying to work with the raw data in assessing this objective indicates the need for an electronic data management system. Such a system would allow reporting of data in a standard format and ongoing assessment using consistent analyses.

Last, lessons learned from examining this objective do not seem to be carried forward to future sampling.

3.0 Detecting and Assessing Regional Trends

The author has equated "cumulative" with "regional", and so the former has been dropped from the title of this objective. In fact, the two terms are not synonymous, and neither has been suitably addressed in this section (see Appendix IV for details).

Objective 2 is broken into three subobjectives: (1) long-term trends, (2) 2000 vs. 2001 comparisons, and (3) upstream-downstream comparisons and trends.

Subobjective 1 – It is hard to imagine why the author tried to identify long-term trends using spotty data from a five-year program. How can identification of long-term trends help a biomonitoring program? How can five years be considered long term, especially when data within the five years are bedevilled by changes in methods and locations and are not consistent?

Guesses as to what is controlling trends seen are pie in the sky; it appears that methods changes are mostly responsible.

Planning for the future is equally chancy; no pilot study or calibration activities are planned beyond letting the sampling run for another five years to see what happens.

The repeated observation that rivers in the study have unique benthic assemblages (p. 6-50/4) is highly dubious (see Objective 1, p. 6-35/4). This statement is hard to believe because of the coarse level of taxonomic identification used in the work, and because rivers in the same region are not likely to have markedly different species of benthic invertebrates in them. However, the author rightly identifies the need for reference rivers, although there is no indication how data from such rivers are to be used (p. 6-50/4).

The author makes no specific recommendation for this subobjective. It is probably a blind alley, and should be dropped as a further goal.

Subobjective 2 – Only the second half of this subobjective comes close to being part of a bona fide biomonitoring program. The power analyses and recommendations that flowed from them (i.e. number of samples, sites sampled, size of samples, etc.) are very useful, and seem to edge toward the EEM program.

The author considers the benthic program is still in its "initial phase" (disappointing because the program has run for five years), so adjusting the sampling design would not entail the loss of an unacceptably large amount of information (p. 6-61/4). The adjustment would also result in better compatibility with historical data. However, why not simply change to an already established biomonitoring program? After all, the author states (p. 6-62/2): "The recommended approach is based on study designs used in pulp mill EEM..." (see also p. 6-60/2). The recommendations from the power analyses seem not to have been used in the RAMP Program Design and Rationale document for future sampling.

The RAMP benthic program could have been further along had information from other programs in the region been used (e.g. EEM, NRBS, NREI). For example, EEM is not sector dependent and the monitoring approach is universal.

Subobjective 3 – It is hard to understand how this consideration adds anything to the program. There is some question about the experimental design used. Upstream/downstream comparisons to measure change are difficult to make if the sites selected are also upstream and downstream of a major tributary. It will never be possible to discriminate between development-related change and tributary effects (in this case, the Christina River is a tributary of the Clearwater River). Thus, the finding that "…existing differences may reflect the influence of the Christina River" (p. 6-63/3) is not surprising.

Results for the Mackay, Muskeg, and Steepbank rivers are also difficult to interpret, especially because the data were collected over three different years. Future experimental design should try to incorporate three types of sites, to evaluate cumulative effects: (1) outside or upstream of all development ("pure" reference sites), (2) downstream of proposed development but upstream of existing development (reference now), and (3) downstream of existing development (affected sites). Spatial comparisons

can then be used to evaluate presence, direction, and magnitude of change to sites either in isolation or as combinations.

Recommendations for alterations of the study design are the same as for Subobjective 2, and derive from power analysis results.

4.0 Monitoring to Verify EIA Predictions

We thought this section would try to use existing data to test the veracity of predictions made by previous EIAs (i.e. are the data collected by RAMP suitable to verify EIA predictions?). Instead, the section is a compilation of EIAs that have been done, with an overlay of benthic monitoring locations. The section considers worthwhile sites, and recommends that less-worthwhile sites be changed. All in all, the section is a paper exercise, rather than being a substantive testing of EIA predictions using RAMP data. Even a compilation of EIA predictions that could be tested using RAMP data in the future would be useful. After five years of monitoring, evaluating the objective by determining whether the data are right to do it – instead of actually doing it – is unsatisfactory.

It is clear that a suitable, overall effects-based monitoring design must be adopted, or development-related change will not be assessed.

5.0 Recommendations and Suggested Implementation

1. Adopt an overall effects-based monitoring program or development-related change will not be assessed. Models are provided by EEM and RCA. In fact, EEM has been operating upstream on the Athabasca River for as long as the RAMP has been around. Adoption of either the EEM or the RCA model would provide the following benefits:
    - the protocols for these programs are well developed, so the details of site selection, sampling, sample processing, and data analysis can be imported directly into RAMP
    - personnel experienced in EEM and RCA are available to offer advice
    - EEM and RCA allow for the addition of sites as oil sands development proceeds
    - reference sites or areas would be included in an EEM or RCA program
    - RAMP could then focus on detecting change rather than on descriptive approaches, and would be able to interpret regional trends and cumulative effects.

    On balance, adoption of the EEM program would be the best choice because it is already operating in the area and because it would cause less disruption than the RCA to RAMP. However, considerable effort will likely be needed to see what elements of RAMP can be salvaged and applied directly to the EEM program.

2. The Athabasca River must be included in any monitoring program for oil sands development. It is the largest, most important ecosystem in the region and will be the receiver of the cumulative effects of development. In the face of EEM,

NRBS, NREI, and PERD programs on the Athabasca River, it is a mystery why RAMP chose to abandon the Athabasca after only one year of study (1997). It should be the core of the RAMP program for all subject areas. RAMP claimed that direct sampling of benthos in the Athabasca River downstream of Fort McMurray was not possible because of shifting substrates. We could not assess this claim because the Five Year Report lacked information and a review of past sampling attempts. RAMP should review industry, research, and provincial benthic biomonitoring programs before attempting other approaches to collect benthos (e.g. artificial substrates) on the Athabasca River.

3. An electronic database management system should be started as soon as possible to enable electronic reporting of raw data in a standard and consistent format and on-going assessment of data using consistent analyses. This recommendation is essential, given the long-term nature of oils sands development. Existing initiatives in Environment Canada's Prairie and Northern Region have integrated provincial and federal water quality data, water quantity data (HYDAT), EEM data for the Athabasca, and point-source quality (i.e. pulp mill and municipal sewage effluents) and quantity data. For example, EcoAtlas-CE has been developed under the NREI program, is currently being expanded to include EIA data, and is available for RAMP to use and develop.

4. The separate components of RAMP need to be better integrated to answer questions and needs between components (e.g. connections between water quality, benthic invertebrates, and fisheries). The overall approach should be an ecosystem-level study, rather than several disparate pieces. The lack of integration amongst aquatic components seriously compromises the ability of RAMP to assess effects-based biological changes.

5. RAMP needs to lean more heavily on regional programs that have been done (e.g. AOSERP, NRBS) or that are underway (e.g. EEM, NREI, PERD) for historical and contemporary information generated and lessons learned. It is also advisable that RAMP activities be more tightly coupled to the CEMA-sponsored Muskeg River study.

6. RAMP has an opportunity to contribute to new functional knowledge, and is encouraged to do so through primary publications. The standard is high for such publications, which means the standard of RAMP activities must also be high.

7. Benthic macroinvertebrates can be used in a variety of ways in biomonitoring activities. RAMP's predominant use has been attributes of community structure (e.g. abundance, density, taxa richness). More use should be made of the biomonitoring potential of benthic macroinvertebrates (e.g. the recent proposal to use mussels as sentinel organisms for contaminants).

ASSESSMENT OF THE FISH POPULATIONS COMPONENT
PREPARED BY JOHN POST, KELLY MUNKITTRICK, MONIQUE DUBÉ AND BRIAN SOUTER

1.0 Introduction

We reviewed mainly the five-year review document, with additional reference to the annual reports produced during the 1997-2001 review period (Post, Munkittrick, Dubé). Souter reviewed the Fish Abnormalities Report as part of the ramp 2000 Annual Report.. Three general objectives are listed in Chapter 7 on Fish Populations: (1) collecting scientifically defensible baseline and historical data to characterize variability, (2) monitoring aquatic environments to detect and assess cumulative effects and regional trends, and (3) collecting data against which predictions contained in environmental impact assessments (EIAs) can be verified.

Three main issues were raised: (1) ensure important fish populations are not aversely affected by development, (2) maintain "ecological integrity", defined as no adverse effects on growth, reproduction and survival, and (3) use early warning indicators. Three additional considerations were raised: i) use statistics to indicate "significant" patterns, ii) use all available data, and iii) link to other RAMP programs.

The review of the Fisheries component had six specific objectives:

1. characterize variability in individual and population-level metrics
2. evaluate program's ability to do (1);
3. identify cumulative effects;
4. evaluate program's ability to do (2);
5. use information collected to verify EIA predictions; and
6. can the program be improved?

This chapter is a summary of the fisheries template reports, which appear in Appendix IV, and to which the reader is directed for greater detail.

2.0 General Comments

The program lacks a clear focus and clear hypotheses regarding what it is trying to do. As it currently stands, the project has suffered from inconsistencies in study design, study area, sampling methods, and quality control practices. The synthesis does not focus on telling us what we should know by now, e.g. what species are resident (in what seasons) and what species migrate here (and when and for how long)? This baseline information is critical to understanding when and how sampling should be conducted.

As it stands the RAMP Fisheries program does not provide a very useful assessment for discerning current impacts or as a benchmark for assessment of future impacts. The collection methods (boat electrofishing) have not been characterized to see what the variability is, and whether they are adequate for the questions (once the questions are developed). The sampling times vary between years, and the synthesis compares fish collected in spring and autumn, resident and non-resident. Much of the statistical analysis is weak or wrong, and does not focus on providing a synthesis that we can use to

move forward. Little attention has been paid to the problems of pseudoreplication inherent in many analyses.

3.0 Characterizing Existing Variability

This section tells us about the species inventories, and only 19 of 30 reported species were seen during the inventory. The species inventory varies because of changes in sites, seasons and sampling methods. The report documents that, for many of the larger species, there are seasonal differences in size of individual fish, suggesting migration into the study area of larger individuals from outside the system. It is crucial to understand fish migration patterns so any effects on fish relative to oil sands developments can b e assessed. Local evaluations have to use fish whose life-history characteristics and performance attributes reflect local conditions. Migrating fish make linkages to development difficult. It is also critical that the surveys use similar sites, methods and seasons, and design the study based on knowledge of the system. The sentinel species should be abundant enough that sufficient samples can be collected, be resident during critical portions of their life cycle, and have measurable characteristics (e.g. if aging is difficult for a particular species than that species may not be a good indicator). Power analysis should be used to ensure that sufficient samples are collected.

The fish tissue analyses are not useful for assessment purposes – PAHs will not accumulate to significant levels in fish muscle until environmental levels are very high. They will be detectable at lower concentrations in bile. The design of the contaminants collections, and study design in general, should be based on hypotheses related to anticipated potential impacts, or specific questions raised by the impact assessments. Furthermore tissue collections were from fish species (whitefish and walleye) that differed from the species collected in the sentinel surveys. In 1998, samples were collected from a reference area. In 2001, samples were collected from only the oil sands area with an n=1. This approach of measuring organics and metals in tissues of different species, from different sites, and in different years, with no replication has no validity.

The sentinel species work is a good first step towards an effects-based program. However, the study design for the sentinel species component needs to be closely evaluated as to its purpose and what questions are being examined. For example, the sculpin component evaluated reproductive development when growth-somatic indices (GSIs) were <2%. Prespawning female slimy sculpin will have a GSI of >35%, so evaluating before their gonadal investment has started does not tell us much about development related changes. In areas where fish cannot be collected between late November and early May, this species may not be a good, potential sentinel for reproductive evaluations. However, if other options for species are limited, there are other potential approaches, including examining the proportion of the population composed of young-of-the-year fish during the early fall as an indicator of reproductive success and recruitment.

There appears to be a lack of understanding of which indicators should be measured in the sentinel surveys and why. In the 2001 Report for example, GSI was measured in slimy sculpin at sites downstream of development on the Steepbank River. The conclusion reached (see comments on the 2001 Report in Appendix IV) illustrate that the

authors do not understand how changes in indicators fit into an overalll effects-based assessment.

Radiotagging studies for the purpose of effects-based assesment should not collect post-spawning fish. Post-spawning aggregations of local and no-local fish, in many cases, represent aggregations from multiple groups of fish that reside in different parts of the river system. If the purpose is to evaluate local impacts, then fish should be collected during the period of suspected maximum residency (for suckers that would be early autumn), and then the fish can be followed. This is especially important in areas like this one, where we know that the seasonal distributions of fish size change, reflecting an influx of large fish at spawning time. The purpose of the study is not to see where fish come from to spawn; it should be to evaluate whether there are local fish, and if changes in olocal fish can be measured relative to development activities.

Difficulties with the counting fence need to be resolved. It can provide very good data.

The fish abnormalities study also falls short as an effort to characterize variability. The report was "cobbled together" from various sources, methods to identify abnormalities were not consistently applied from year to year, and reporting was inconsistent. There is also no photographic record provided to support result interpretation. There were no links made between water quality and the growths and lesions observed.

4.0 Detecting and Assessing Regional Trends

Much of the field sampling involved inconsistencies in methods and spatial and temporal coverage, rendering the pattern analysis biologically uninformative. A more focused, hypothesis driven, mechanism-based program would be more efficient and likely much more informative in the long run. It is necessary to standardize sampling sites and methods to allow the proper assessment of trends.

There appears to be confusion on the linkages between species selected to characterize variability and species selected to measure development-based change. Monitoring suckers during spawning runs and in the absence of a suitable reference site confound any interpretation of change. Measuring tracers in a different set of species confounds the issue further. If the goal is to measure changes in fish due to oil sands activities then select a resident sentinel, select a reference site(s) (see Appendix IV) and select a tracer for that sentinel. The work by J. Parrott, NWRI, Burlington, Ontario on the Steepbank River is a good example of how this can be done.

Several statements are made in the Five Year Report and Annual Reports indicating that changes measured in sentinels might be due to natural factors. Parrott does an excellent job illustrating a study design that separates a reference site from a site exposed to natural oil sands seepage and from a site downstream of development. In this instance, changes in EROD and sex steroid activities in a sentinel showed clear spatial changes. The importance of reference or "lo-impact" sites to separate natural changes from man-made disturbances cannot be emphasized enough.

5.0 Monitoring to Verify EIA Predictions

EIA predictions were divided into fish habitat, species composition, abundance, health and tissue tainting. It was recognized in the report that habitat was limited in the first two years and discontinued. It was also recognized that the inventory and abundance data was restricted, sites varied, and the selected method was size-selective. The fish health component is a recent addition to the program and, as mentioned above, there may need to be some consideration of the study design for this component. Tissue tainting studies have been conducted, but it is important to separate the questions of tainting, contamination, and violation of EIA predictions (i.e. PAH accumulation).

There are no EIA predictions included in the report. Several generic fish health characteristics are listed but they are not associated with impact predictions. Therefor, it is not possible to assess how RAMP could be used to test EIA predictions.

6.0 Conclusions

If the study is going to use monitoring to tell us something, it has to accept that such monitoring needs to use state-of-the-art technology, needs to be science-based, needs to focused on adaptive management, and has to be committed to telling us about the variability and confidence we can place in conclusions. The main objective should be to initially document, for specific reaches of river, representative reaches and regional reference areas, what species use the area, when they use it, why they use it, and how variable it is. Once these data are available, the baseline monitoring program needs to be developed, using specific questions focused on what the expected changes would be, and what the specific monitoring objectives are. If the study wants to do this, it should commit the money to do it properly, to regularly evaluate progress, have an external science advisory committee, and commit to science-based development of the information needed.

The objective to recognize and incorporate Traditional Ecological Knowledge (TEK) into the monitoring and assessment activities might be considered relevant to the fish abnormalities component. Fish abnormalities are a concern to First Nations in the area and we had expected that a report on fish abnormalities would have some reference top TEK.

ASSESSMENT OF AQUATIC VEGETATION COMPONENT
REPARED BY MARLEY WAISER

1.0 Introduction

The OIL Sands Regional Aquatics Monitoring Program (RAMP) was initiated in response to the large increase in oil sands mining and related developments north of Fort McMurray and the need to coordinate environmental monitoring activities so that potential cumulative effects could be identified and addressed. RAMP was initiated by Suncor Energy Ltd., Oil Sands, Syncrude Canada Ltd., and Shell Canada Ltd.

Surveys of wetland vegetation were conducted in 1997, 1998, and 2001 as part of the RAMP program. Three wetlands sites, Shipyard (adjacent to Suncor's Steepbank Mine), and Kearl and Isadore's lakes (adjacent to Shell's proposed Muskeg Mine Project), were sampled in each of these years. During this time, an effort was made to find a suitable reference wetland site. In 1997, the reference site was Lease 25 wetlands but this site was dropped in 1998 either due to poor access or because it was too close to future oil sands development. In 1998, Spruce Pond was investigated as a possible reference site but it too was dropped due to its hypertrophic status, which made comparison to the other less-enriched sites impossible. In 2002, McClelland Lake was chosen for sampling, although it is unclear from the material provided why this site was chosen.

Vegetation was documented by: mapping wetland classes according to the Alberta Wetland Inventory and using aerial photographs; photographing vegetation from fixed points; conducting a vegetation survey along fixed transects (evaluate species present and relative percent cover); recording vegetation vigor and health; and collecting water quality parameters (water depth, pH, conductivity, dissolved oxygen, percent dissolved oxygen, total dissolved solids and temperature). In 2001, the program was expanded to include calculation of species richness, species diversity (Shannon-Wiener), an index of similarity (Jaccard's), an index of dissimilarity (Bray-Curtis) and some limited statistical analyses (Kruskal-Wallis nonparametric tests).

2.0 Characterizing Existing Variability

The 1997 yearly report states that the objective of the wetland vegetation program, was to provide a description of wetland types, plant species composition and vegetation health as a baseline for future monitoring. In 1998, the scope of the study was to further describe the vegetation communities in Isadore's, Kearl and Shipyard lakes (second year of data to describe natural variability) and to identify and evaluate reference wetlands. In 2001, the stated objective was to continue the task of characterizing the natural variability in the wetland types representative of the three study lakes. The 2001 report also states that the key to RAMP success is to select and verify monitoring methods that will differentiate effects of oil sands development from natural variability and existing anthropogenic effects. The existing reports for 1997, 1998 and 2001 have done a good job, but only of describing the wetland types, plant species composition and vegetation health. The reviewer, however, could not find a clear statement in any document provided of which monitoring methods RAMP investigators selected and verified to differentiate natural from anthropogenic variability.

Wetlands are highly variable ecosystems and teasing out anthropogenic variation from that which is natural is not an easy task. The reviewer is concerned that the sampling frequency (once but at the most twice per year) is too low. Consequently, researchers may not be able to distinguish natural variability from that which may be anthropogenic or arise from the effects of oil sands development. In the 2002 document, under sampling frequency, no mention is made regarding the number of times per year wetland vegetation will be sampled. If sampling can only be done once per year, then it should be done when the plant community is at its peak and at a time of year when the greatest impact from oil sands development is expected (i.e. the time of year when problems are most likely to occur – usually called the index period). According to the US Environmental Protection Agency (EPA) (EPA, 2002 - #4 Study Design for Monitoring Wetlands http://www.epa.gov/waterscience/criteria/wetlands/), "once wetland condition has been characterized, one-time annual sampling during the appropriate index period may be enough for multiple-year monitoring of indicators of biotic integrity. However, metrics and ecological indicator development [which as far as the reviewer can tell have not yet been established by RAMP], may require more frequent sampling to define conditions that relate to the stressor or the impact of interest". Sampling frequency must be addressed by the investigators and this must be done before the next wetland vegetation monitoring takes place.

One of the clear objectives of RAMP is to characterize variation. Because of the high natural variation associated with wetlands and the fact that RAMP is supposed to be investigating the effects of oil sands development on aquatic ecosystems, it is imperative that reference site(s) be found. The current lack of a reference site precludes the ability to detect what is natural variation and what is anthropogenic. According to Richardson and Vymazal (2001) "Reference or undisturbed areas must be included in all biomonitoring analyses if changes in communities are to be assessed accurately". Finding and sampling a large number of reference sites to define regional variability may not be necessary if physically similar sites (size, hydrology, elevation, etc.) can be found close to the disturbance site but out of range of possible disturbances. Reference sites such as these should be selected based on physical or chemical attributes not affected by human intervention (elevation for example) (Rader and Shiozawa, 2001). If a few local reference sites cannot be found, there are other options. For example, sampling a number of minimally affected sites could work (Wright et al., 1995). As well, there is the possibility of establishing reference conditions within each wetland using paleolimnological techniques, providing that sediments have been relatively undisturbed through time. Finding reference sites is a "must do" for the RAMP program. Although the 2002 report does identify the need for reference sites, it should be at the very top of the list of what must be done in order for RAMP to become an effective scientific program that fulfils its objectives.

3.0 Detection of Effects and Monitoring Cumulative Effects

In the Executive Summary of 1997, a statement is made that RAMP is largely an effects-oriented project whose priority is early detection of potential effects. The stated objective of the wetland vegetation subprogram is to "provide a description of wetland types, plant

species and composition and vegetation health as a baseline for future monitoring". This objective is not in line with the stated objectives of the overall program; the emphasis for this subprogram should be on effects monitoring with regard to wetland vegetation, not monitoring of vegetation. If the objective is on effects, then the rationale has to be more clearly defined.

The existing rationale is stated as follows: "wetland vegetation has been documented as an important biomonitoring parameter for examining potential effects to wetland systems". But that is where it ends. The reviewer agrees with this statement, but unfortunately, the investigators do not seem to have thought about how they are going to use the data that they have collected to demonstrate effects of oil sands development. There is no clear, well thought-out, scientifically based plan in place detailing why all of these data are being collected and how these data will be used to detect effects. Nowhere in the documentation could the reviewer find a clear statement of what constitutes unacceptable change in wetland vegetation. Consequently, RAMP has failed miserably with respect to meeting the objective of detecting effects and monitoring cumulative effects.

As a first step, the investigators need to figure out which aspects (attributes = measurable components of a biological system) of wetland vegetation are the most likely to respond to disturbance resulting from oil sands development. Karr and Chu (1999) point out that "a bewildering variety of biological attributes can be measured but only a few provide useful signals about the impact of human activities". Consequently the careful choice of attributes, which will show a consistent response to oil sands development, is imperative. The goal would be to identify those vegetation attributes that respond reliably to human activities, are minimally affected by natural variability, and are cost-effective to measure (EPA, 2002; #6). The data have probably already been collected, so it would be a matter of sorting out which of the vegetation parameters measured are most likely to respond to the stressors provided by oil sands development.

Attributes that respond to human disturbance are called "metrics". Metrics can be divided into three groups: community based, metrics based on plant functional groups and species-specific metrics (EPA, 2002). Metrics are used to detect ecological impairment and diagnose causes of impairment. This approach has been widely used in wetland research. In a study of 26 Minnesota wetlands, for example, an index of biological integrity was developed using 10 vegetation metrics (Helgen and Gernes, 2001), in an effort to compare the biological integrity of reference wetlands to wetlands in agricultural areas or those receiving stormwater inputs. Vegetation metrics included the number of vascular genera, number of nonvascular taxa, sum of all sedge species cover classes, sensitive species, tolerant taxa, grass-like taxa, monocarpic species, number of aquatic guild species, distribution of cover in a sample and sum of persistent litter taxa-cover classes. Scoring criteria were developed by sorting metric values from high to low and then dividing the data into three groups. The maximum score for the index was 50, whereas the minimum was 10. A reference wetland in a state park received a score of 50, whereas one agriculturally affected site received a score of 10 (Helgen and Gernes, 2001). Such an approach would have great applicability to the

RAMP study (refer to the EPA website http://www.epa.gov/waterscience/criteria/wetlands/ and the module "Using vegetation to assess environmental conditions in wetlands" for an in-depth discussion of the use of metrics for wetland evaluations).

4.0 Monitoring to Verify EIA Predictions

The reviewer could not find reference to "monitoring to verify EIA predictions" in the documents provided concerning wetland vegetation. If this is one of the objectives of RAMP, then this oversight needs to be addressed.

5.0 Recommendations and Suggested Implementation

1. Change objectives and rationale so that they are clearly stated and scientifically based. The investigators need to look at the monitoring program and decide, based on an intensive search of the scientific literature, which attributes of wetland vegetation they should be monitoring, i.e. which attributes will give the most information regarding variability (natural and anthropogenic) and effects of oil sands development.
2. Work done must reflect the objectives and rationale. The investigators need to keep their focus on what the objectives of the research are and make sure that the research they propose will meet the stated objectives. To date this has not been done.
3. The time of year of sampling and sampling frequency for wetland vegetation needs to be re-examined. If sampling can only be done once per year, then it should be done when the plant community is at its peak and at a time of year when the greatest impact from oil sands development is expected (i.e. the time of year when problems are most likely to occur, usually called the index period). According to the EPA (2002 - #4 Study Design for Monitoring Wetlands http://www.epa.gov/waterscience/criteria/wetlands/), "once wetland condition has been characterized, one-time annual sampling during the appropriate index period may be enough for multiple-year monitoring of indicators of biotic integrity. However, metric and ecological indicator development [which as far as the reviewer can tell have not yet been established by RAMP] may require more frequent sampling to define conditions that relate to the stressor or the impact of interest". Due to the high variation within wetland systems, if one is going to compare systems then it is important that sampling be done at the same time of the year on a year-to-year basis. According to EPA (2002), "the establishment of a standard sampling window ensures that representative results are obtained at each site and that valid comparisons can be made between different wetlands". Wetland vegetation sampling was not done in the same month from year to year in the RAMP study. This should be addressed for future sampling efforts.
4. Establish attributes of wetland vegetation that are metrics, i.e. attributes that are appropriate for monitoring the effects of oils sands development on wetland vegetation (see comments in Appendix IV). Then base the wetland vegetation

monitoring program on measuring those metrics. Combine the metrics into a multimetric index which will allow the investigators to score and compare affected sites to reference sites. In this way, effects can be measured. Karr and Chu (1999) point out that "a bewildering variety of biological attributes can be measured but only a few provide useful signals about the impact of human activities". Consequently the careful choice of attributes that will show a consistent response to oil sands development is imperative. The goal would be to identify those vegetation attributes that respond reliably to human activities, are minimally affected by natural variability, and are cost-effective to measure (EPA, 2002; #6). The data have probably already been collected, so it would be a matter of sorting out which of the vegetation parameters measured are most likely to respond to the stressors provided by oil sands development.

5. Because of the high natural variation associated with wetlands, and to meet the stated objective of determining effects, it is imperative that a reference site or sites be found. This must be done if researchers are to meet the objective of determining the effects of oil sands development on wetland vegetation. Without a reference site, collection of more vegetation data would be a waste of time and effort. The current lack of a reference site precludes the ability to detect what is natural variation and what is anthropogenic (due to oil sands development). According to Richardson and Vymazal (2001) "Reference or undisturbed areas must be included in all biomonitoring analyses if changes in communities are to be assessed accurately". Reference sites serve as the standard against which other sites will be judged. Finding and sampling a large number of reference sites to define regional variability may not be necessary if physically similar sites (size, hydrology, elevation, etc.) can be found close to the disturbance site but out of range of possible disturbances. Such reference sites should be selected based on physical or chemical attributes not affected by human intervention (Rader and Shiozawa, 2001). If a few local reference sites cannot be found, there are other options. For example, sampling a number of minimally affected sites could work (Wright et al.,1995).
6. Consult with a statistician to improve not only the way that data are analyzed but also how to better integrate the vegetation data with the water chemistry and quality data.
7. Improve referencing to scientific literature – don't base your study solely on technical and government reports.
8. Don't wait five years for a review. Have an outside objective scientific panel with the appropriate experience and expertise review work done on a yearly basis.
9. Less representation by industry and more representation by non-partisan groups (Environment Canada, universities, etc.) is advised. The make-up of the RAMP committees is too heavily weighted towards industry. The lack of scientific expertise on these committees is reflected in lack of scientific rigor in the RAMP wetland vegetation reports.
10. Proposed research should be vetted first by outside experts. Before going out into the field to collect data, submit proposed monitoring and effects research to appropriate qualified scientific personnel for review and comment.

ASSESSMENT OF ACID SENSITIVE LAKES COMPONENT
PREPARED BY MICHAEL TURNER AND JAN BARICA

1.0 Introduction

The Acid Sensitive Lakes (ASL) program was designed to provide an early warning of the effects of acid deposition emanating from the Oil Sands Region. In particular, a properly designed ASL program will support the Cumulative Environmental Management Association (CEMA) objective of activating "the management response in the case of a yellow or red condition is intended … to ensure there are no exceedances of management objectives beyond the level of protection area". The ASL component of the Oil Sands Regional Aquatics Monitoring Program (RAMP) was initiated in 1999 in partnership with Alberta Environment. This review was carried out by two individuals with two complementary backgrounds. The senior reviewer has considerable expertise on the effects of acidification on lakes, while the second reviewer has broad experience in water quality monitoring systems in lakes and rivers in many areas of the world (see Appendix II).

Although there are no trends yet seen that support the idea that acidification is occurring, the ASL program reports that there are "already some concerns regarding acidification in the Oil Sands Region in the foreseeable future".

The potential for acidification is of concern because acidification represents one of the most seriously damaging impacts that humans can wreak upon ecosystems. Impacts can range from the physicochemical to biological changes that alter the structure and function of these ecosystems. Biological changes include irreversible impacts upon habitat productivity, foodweb integrity, ecosystem health and biodiversity.

An objective of the CEMA framework was to "avoid change in water chemistry that will result in change to ecological receptors either in the short term or through a long-term trend". This objective stemmed from the recognition that "it is possible that some change in water chemistry will occur from anthropogenic emissions. Any such change will be limited so that it is consistent with the management framework goal".

Selection of Lakes

Up to 50 moderately to highly acid sensitive lakes in northeastern Alberta (i.e. the region expected to be impacted by Oil Sands development) have been selected for regular monitoring, although this number has varied from year to year. In 2002, 39 lakes in the Oil Sands Region were included to represent a gradient in acid deposition; also we used 5 lakes in the nearby Caribou Mountains plus 5 lakes in the Canadian Shield that are distant from sources of acidifying emissions (reference lakes). These lakes were deemed to represent systems that were moderately to highly sensitive to acidification (<20 mg/L $CaCO_3$), close and away from the Oil Sands area, and accessible by at least float plane.

Table 10.1 (2002 Report) presents modeled acid deposition rates with critical loads calculated for individual RAMP lakes (developed by CASA-established guidelines in 1996-1999). The critical load is defined as the highest load that will not cause chemical

changes leading to long-term harmful effects on the most sensitive ecological systems (study done outside RAMP), set at 0.25 keq $ha^{-1}yr^{-1}$ for sensitive soils in Alberta, taking into account the expected buffering capacity of the lakes and input of base cations for the watershed. It represents the amount of acid deposition below which acid neutralizing capacity (ANC) or pH remain above a specific threshold value (ANC set at <5 ueq·L^{-1} or pH 6 for the Oil Sands Region by the NO_x – SO_x Management Working Group (outside RAMP).

Sampling Program
The lakes have been monitored annually. Vertically integrated collections from the euphotic zone from up to five sites in each lake were combined to form a single composite sample for chemical analysis. Full vertical profiles of Secchi depth, dissolved oxygen (DO), temperature (T), conductivity and pH are done at the deepest location in each lake. Parameters monitored include standard (routine, generic, core) indicators used in water quality monitoring, both of acid lakes and other water bodies (i.e. pH, total suspended solids [TSS], total dissolved solids [TDS], alkalinity, bicarbonate and other major ions, nutrients, DO, etc.). Table 8.2 of the same report presents a detailed rationale for core ASL monitoring program, with general guiding principles, site selection, and specific methods and procedures.

Scope of Material
The materials considered for this review included: RAMP 1999: 2.1.4, 3.1, 8.1, 9.1.4, 9.2 RAMP 2000: 2.1.4, 4.4, 8, 10.1.4; RAMP 2001: 2.1.4, 3.5, 4.2.4, 10, 11.1.3, 12.1.4 RAMP 2002: 1.2.6, 2.1.5, 3.6, 4.2.6, 10; RAMP Five Year 1997-2001 Report (ASL sections 1.4.5.2, 1.5.5.3 and 1.6.5.3); RAMP Program Design and Rationale (2002) section 8 and Table 8.1; Horizon Oil Sands Project Application (technological aspects); and notes on the RAMP 2003 Oct. 22 meeting. As well, the reviewers examined supplementary material provided by B. Kemper including: CEMA: Acid deposition management framework recommendations for the oil sands region of north-eastern Alberta; CEMA Research priorities and monitoring enhancements related to acidification and the management of critical loads in north eastern Alberta; and Preliminary review of the effects of acid deposition on northern Saskatchewan lakes (D. Ballagh, 1999).

It is important to recognize that the lack of an integrated overview document (as was available for other RAMP projects) diminished the effectiveness of the review and significantly increased the effort required. Frequent changes to the objectives and scope of the review further diminished the effectiveness of the review planning, and have arguably caused the review to be incomplete. It is imperative that a consolidated report for the ASL program be prepared that includes documentation of the linkages with other programs (e.g. in diagrammatic form). Only then can an effective review of the program be conducted.

2.0 Assessment of Acid Sensitive Lakes Program

The program as it has been described in the annual reports is unlikely to achieve its stated objectives, although implementation of the several studies recommended by the CEMA NO_x - SO_x working group would markedly improve the program and its effectiveness. If

the CEMA recommendations are not acted upon, it is unclear that the program can achieve its stated objectives of:

1. Collecting scientifically defensible baseline and historical data to characterize variability in the acid sensitive lakes;
2. Monitoring aquatic environments to detect and assess cumulative effects and regional trends; and
3. Collecting data against which predictions contained in environmental impact assessments (EIAs) can be verified.

Assessment related to ASL Program Objectives

The ASL program's first objective is to collect scientifically defensible baseline and historical data to characterize variability in the oil sands area. Despite the limited frequency of sampling (once a year) and short length of the monitoring program (1999-2002), the Program has delivered some useful information and new knowledge. However, it could not collect scientifically justifiable baseline and historical data to enable characterization of the variability in the Oil Sands Region. Nor was a procedure proposed that would enable valid statistical detection of trends in the future; currently, data are insufficient for a trend analysis.

The second program objective was to monitor aquatic environments in the Oil Sands Region to detect and assess cumulative effects and regional trends. The data collected since 1999 have been insufficient to detect any regional trends or cumulative effects of acid depositions. (Nor would we expect to detect a trend in four years.) However, sulphate concentrations in several lakes of the Birch Mountains in the Oil Sands Region are already high, and are similar to or exceed values seen in experimentally acidified lakes of the Experimental Lakes Area (ELA) at their most acid. This suggests that some acidification may already have occurred. This observation applies only to the chemical parameters monitored because so far there is no biological monitoring being conducted in the ASL program.

The third objective was to collect data against which predictions contained in environmental impact assessments (EIA) can be verified. The power of the monitoring program described in the RAMP ASL program annual reports is insufficient to verify EIA predictions. Principle concerns include:

- sampling frequency is inadequate to monitor parameters that are known to be seasonally variable;
- the timing of sampling avoids possible spring acid pulses that occur elsewhere in acid impacted regions;
- many important early warning lake responses are biotic and these are not being monitored;
- some of the lakes being monitored are not particularly acid sensitive; and
- deposition (including dry deposition) is not being monitored forcing decisions to rely solely on modeled scenarios.

See below for additional concerns.

Assessment Related to Linkages and Integration

Within the Program: The water chemistry of the acid sensitive lakes appears to exist largely in isolation of other components of RAMP; certainly other components were excluded from the ASL reports. Although some phytoplankton and zooplankton samples have been collected, no plans have yet been identified to have them analyzed or interpreted.

RAMP to Region: Although the selection of sampling stations seems acceptable, it is unclear how representative the selected lakes are from a regional perspective. Certainly it is good that more than one cluster of lakes is being studied. However, there are good reasons for adding at least one or two more clusters, including the acid sensitive lakes in northwestern Saskatchewan.

RAMP to Other Programs: It is also unclear from the annual RAMP ASL reports what is going on in other (possibly related) ASL studies in northern Alberta because the annual reports have been presented largely in isolation of such activities. There is no evidence of any linkage of this component to other environmental monitoring programs in the annual reports except for Table 8.2 of the 2002 RAMP Program Design and Rationale. Recently received information from Bryan Kemper indicates that there are several important additional efforts proposed by CEMA that are outside the activities identified in the RAMP annual reports. These proposed studies and their interactions with RAMP monitoring need to be linked in a summary report.

Concerns and Gaps

The ASL Program provided useful and scientifically valid information that will contribute to regional, national and international understanding of relationships of various components of northeastern Alberta acid sensitive ecosystems. There have been some adaptive changes made to the program, although sometimes the changes have not always been well implemented (e.g. although gran alkalinity began to be measured in the second year, the older measurements remain the reported values). However, there are so many serious issues that remain to be adapted to that we believe the experimental design described in the ASL Program RAMP reports is unsuitable for testing the program objectives.

We are concerned that the gaps in the present ASL monitoring program will prevent development of a statistically sound base to assess the variability of the selected parameters and to develop even an indication of acidification trends. These gaps include:

1. Inadequate sampling frequency and inappropriate timing of sample collection: once-a-year sampling is insufficient. A single annual sample of water chemistry collected from each lake cannot provide an adequate assessment of the average values of any chemical substance that is nonconservative (i.e. most of those that are of interest such as pH). Given the shallow nature of many of these lakes, and probable rapid water renewal, it is likely that water chemistry conditions in the lakes are highly variable. For example, there are cases where interannual differences in pH exceed one unit even after only three years of sampling, but

probably not as a result of changes in acidic deposition. As a result, the power to detect interannual differences or trends in response to changes in loading of acidic substances will probably be exceedingly low.

The spring acid pulse has been neglected, yet it may indicate the impact of accumulated deposition over the winter. Program changes in lake selection will likely make it more difficult to detect temporal trends.

2. Biological indicators are missing that could better and more sensitively identify the effects of acidification on aquatic ecosystems. In addition to the chemical characteristics of ASL, we need to look at their biota, and functional and structural indicators, such as those that are related to productivity and biodiversity. Other biotic indicators include changes in phytoplankton species composition coupled with shifts to acidophilic genera, changes in zooplankton assemblages, and altered phytobenthic, zoobenthic and fish productivity. Such indicators have been useful elsewhere in the study of acid sensitive lakes world-wide, and would yield a more convincing demonstration of the effects of acidification on aquatic biota, which should be our primary concern.

3. Often metals in addition to acidity per se can be biologically damaging. The sub-program ignores measurement of any metals (e.g. mercury and aluminum) even though metals are sampled and analyzed in other subprograms.

4. It is unclear how changes are to be detected in the monitored lakes. The lack of a scientifically challengeable hypothesis prevents objective evaluation of the monitoring data in order to detect temporal trends. It is also unclear what quantitative criteria will be used for detecting change. There is discussion of several acid sensitive parameters (e.g. alkalinity [gran or fixed-point titration] or ratio of bicarbonate:divalent cations). There are also analyses of year-to-year trends using several crude means (eyeballing clustered histograms or box and whisker plots of pH and alkalinity). Yet there is no statement of what parameter, rate and degree of change or technique of analysis will be used to assess whether acidification is occurring. Also is acidification to be evaluated on a lake-by-lake basis, or as a result of a regional cluster analysis?

   Detection of trends in the monitored lakes will be challenging because of fluctuations in the sampling program. Although the program has been adaptive in some respects, i.e. adjusting methods and lakes, such adjustments can increase the difficulty of detecting long-term trends. For example lake selection varied over 1999-2002 in a relatively nonsystematic way; 38 lakes were sampled (Table 3.27 of 2001 report), although only 27 were sampled in all 3 years. What precautions will be taken to factor out the influence of sampling irregularity on the ability to detect temporal trends? Moreover, the criteria for lake removal and addition are sometimes unclear. For example, in the Oil Sands Region, the pH was higher in the replacement lakes (A300, L29) than in the lakes dropped (A47, L1, L30) (compare pH in Figure 10.1 of RAMP 2001 Vol. 1). It would seem that selecting

higher pH and higher alkalinity lakes runs counter to the principle of selecting acid sensitive lakes.

5. It is noteworthy that the acid deposition rates in the Oil Sands Region were modeled, rather than measured, in a number of recent EIAs for oil sands development (six companies listed). It appears that there is no verification step to ensure that the lakes are actually receiving the modelled acidic inputs. The primary focus was on modeling the Potential Acidic Input (PAI) in keq $ha^{-1} \cdot yr^{-1}$, including wet and dry deposition by sulphur and nitrogen compounds from sources within the area and from background sources, accounting for the mitigating effect of base cations (Table 10.1, 2001 Report). PAI values are expected to represent potential "near-future" deposition rates, as some yet undeveloped (i.e. planned and/or approved) projects were considered in modeling. But no depositional data are provided in this section to substantiate that these PAI are likely to be correct. As a result, the lack of verifiable depositional information diminishes the validity of future projections, increases the uncertainty of interpreting the monitoring observations, and limits the ability to evaluate the responsiveness of the monitored lakes.

   Furthermore, the CEMA document identifies the Henriksen model as "difficult to apply and validate in low-relief wetland-rich terrain". We concur that there is need for a dynamic model that is adapted to the northeastern Alberta region, and that has been verified. The CEMA report identifies efforts that could result in model development and verification.

6. The lack of hydrology and chemical data for the watersheds of the study systems limits the understanding of the relationship between the aquatic chemistry data being collected and the acidic deposition that is occurring.

   Because the lakes selected in the Oil Sands Region are predominantly shallow, they are likely to have relatively rapid water renewal times. (Although these data are not presented, a hydrologist could provide theoretical water renewal rates based on average catchment hydrological yields, average precipitation, and photometric assessment of catchment areas.) As a result, many of the monitored lakes will likely reflect terrestrial catchment influences more strongly than in-lake processes, which would predominate with longer water renewal times. Therefore, it is less clear how the ASL program will serve as an early warning of excess acid deposition.

   The lakes selected may be relatively insensitive to changes in acid loading for yet another reason. Lake trophic status could confound the ability to detect acidification because only one lake is oligotrophic, and the rest range from mesotrophic to hypereutrophic status. Typically oligotrophic lakes are more acid sensitive than are eutrophic lakes, which can have greater acid buffering capacity (e.g. ELA's L302N experiment evaluating nutrient additions on alkalinity

generation, and P. Dillon's similar experiments in the Dorset region of southeastern Ontario).

7. Conventional measurements may be insufficient to characterize the chemistry of the monitored lakes, which often have high concentrations of dissolved organic materials. Additional information is required about the buffering capacity in such aquatic ecosystems of the organic complexes that are common in the lakes of the Oil Sands Region.

8. The monitoring program does not distinguish between acidic emissions from the oil sands operations from other regional or long-distance sources. Perhaps there is some marker of the oil sands operations that will enable oil sands emissions to be distinguished in the depositional areas from background deposition or from other sources. Routine parameters such as pH, alkalinity, N- and S-compounds, and base cation ratios are so far the only parameters used in monitoring of acid sensitive ecosystems world-wide. Although this is a weakness of all ASL monitoring programs, in the event of increased deposition, it will be difficult to identify the source of acid emissions.

9. The idea of including "reference sites" or lake clusters in the ASL program is excellent. However, it is unclear what criteria were used for selecting these reference sites. How have these reference sites been matched with the Oil Sands Region lakes? It is also unclear how the reference data will be used to assess temporal trends in oil sands emissions-affected lakes.

3.0 Recommendations and Suggested Implementation

Independently of our review, the CEMA report "Research Priorities and Monitoring Enhancements …" made several recommendations that are germane to the objectives of the RAMP ASL program. In many cases, the recommendations pertain to issues of concern that we have identified in our review and, as a result, overlap to some degree with our recommendations. As such, these projects merit mention, and we encourage that they be considered for incorporation into, or refinement of, the ASL program. The germane projects that the CEMA report has recommended include:

- early detection of acidification of small watersheds and dynamic model development;
- hydrologic regime of potentially acid sensitive lakes – determining annual through-put flux;
- determining the mechanism of organic acid buffering and its response to anthropogenic deposition of sulphur and nitrogen;
- seasonal changes in lake chemistry;
- determining historical changes in lake chemistry and relationship to productivity using paleolimnology; and
- coupling the ASL program with other relevant model verification and terrestrial monitoring studies.

The primary recommendations that we suggest for improving the ASL program are in order of priority:

1. Integration of programs and research plans
   a. Integrate the RAMP-related ASL program with other programs, e.g. the CEMA NO_x-SO_x Working Group-related efforts. A matrix of activities, organizations and their linkages needs to be presented that defines well the context of RAMP's ASL program; Table 8-1 of the 2002 RAMP Program Design and Rationale is an incomplete start. Although integration of RAMP's ASL program with other efforts may already be underway, without a summary report it is unclear that this is so. If the ASL program is actually a separate endeavour from these other activities, then substantial efforts are needed to unify these ASL-related activities to avoid "reinventing the wheel", and wasting resources.
   b. A related recommendation is that there should be a coherent and integrated monitoring and research plan put forward. Exclusion of the ASL program from the final report was incorrect. In the absence of a final report for the ASL program, there is little evidence of a plan for 2004-2009 except for continuation of monitoring efforts. If we have to project forward what we have seen through 2002, then the plan cannot be considered satisfactory. Note that some of what could be a plan for 2004-2009 appears to be embodied in CEMA documents.
   c. Coupled with these coordination efforts is the need to ensure that the ASL program is well linked to regional monitoring of the deposition of acidic substances. This monitoring must also include monitoring of dry deposition, which recent information from Environment Canada (Bob Vet) indicates could be a large component of total deposition (ca. 30-50%). Reliance on unverified modeled deposition is unsatisfactory.

2. Proposed changes to the current monitoring program
   a. Clearly state the working hypothesis or question that is to be tested in detecting long-term changes in acid status of the monitored lakes. State the criteria that will be used to test that hypothesis.
   b. Increase the sampling frequency within each year using an analysis of the power to detect change, and adjusting the sampling effort accordingly. (CEMA notes that sampling for the US Environmental Protection Agency [EPA] monitoring program occurs four or five times a year.)
   c. Introduce spring-time sampling as a priority. Exclusion of spring-time samples precludes the ability to detect acid pulses in the monitored lakes.

3. Additional parameters to be introduced into or integrated with the RAMP monitoring program
   a. Add an in-lake biological component to this study. This would both help with the evaluation of the biotic sensitivity of the systems, and enhance the power to detect change. Relatively inexpensive possibilities include phytoplankton

and zooplankton; currently planktonic samples are collected but there is neither a plan nor resources for their analysis and interpretation. More energy intensive alternatives include study of fish populations, zoobenthos and benthic algal assemblages. A further expansion would be to consider waterfowl usage of these systems such as is done by Environment Canada's Canadian Wildlife Service.

b. Add a metal component to the program (perhaps linking to the water quality component), particularly aluminum and mercury. For example it has often been reported that mercury bioaccumulation can be increased in acidifying systems. Mercury contamination can be serious for the health of wildlife, for domestic fisheries and for recreational fisheries. It is likely that the oil sands emissions will also include increased deposition of mercury in the downwind regions. Hence, mercury could be increasing in aquatic biota both because of increased deposition and because of pH-related changes.
c. Several of the monitored systems need to be better characterized in terms of their watershed characteristics, including their lake bathymetry and rates of water renewal, for example; CEMA has made a similar recommendation.

4. Proposed research needs to complement the RAMP monitoring program
   a. Establish intensive study watersheds that are known to be acid sensitive and are receiving acidic inputs. These sites should be hydrologically calibrated, and information should be gathered that defines well the biological and chemical properties of the lakes in the context of their watersheds and depositional regimes. (Note that CEMA shows this as a proposed study.) Extra effort directed to these systems would be designed to help interpret the broader regional results.
   b. Spend effort to understand the role of organics in the acidification and buffering of these lakes. (This has also been recommended by CEMA.)

5. Suggested modifications to the lake selection
   a. Once depositional information is available, it should be verified that the lake cluster deemed to be a suite of reference lakes is actually suitable for this purpose.
   b. Add downwind lakes in Saskatchewan that are known to be acid sensitive, known to be receiving acid deposition, and projected to acidify.

OVERALL ASSESSMENT OF THE RAMP AND RECOMMENDATIONS FOR THE FUTURE

Introduction

The Oil Sands Regional Aquatics Monitoring Program (RAMP) in the Oil Sands Region of northeastern Alberta was designed to measure baseline environmental conditions, and predict and assess effects from proposed developments. RAMP was designed as a long-term monitoring program that incorporates both traditional and scientific knowledge. This review has focused on three major objectives of RAMP, specifically: (1) characterizing existing variability, (2) detecting regional trends and cumulative effects, and (3) monitoring to verify environmental impact assessment (EIA) predictions. Following the organization of the program and the annual reports and the Five Year Report, the review was divided into seven components viz., climate and hydrology, water quality, sediment quality, benthic invertebrates, fish populations, aquatic vegetation and acid sensitive lakes. Our overall assessment is based on the narrative reports found in the previous sections, the template-based reviews (Appendix IV) and separate discussions with some of the component reviewers. In this section, we present a number of issues and concerns that were common to several different components and program objectives. Based on the assessments, we make recommendations for future action. Our recommendations are separated into three types: (1) organizational, (2) primary technical, and (3) secondary technical.

We saw many signs of positive progress with RAMP. The very existence of a major regional aquatic monitoring program is a positive sign for Alberta. Initiating joint monitoring by the oil industry in 1997 was a progressive initiative leading to benefits now and in the future. The companies involved are to be commended for their vision and their significant financial contribution over the years. A long-term initiative such as RAMP is rare.

The RAMP initiative to draw individual components into a comprehensive regional aquatic monitoring program is a positive step towards relevance and effectiveness. This is a major region of Alberta and is an area of significant environmental disturbance. RAMP offers an important opportunity to ensure environmental protection, support environmental rehabilitation in the future and enhance our level of knowledge and understanding of boreal aquatic ecosystems in disturbed and undisturbed settings.

The general consensus of the reviewers was that the Five Year Report was well organized and written in a manner that is accessible to most stakeholders, with a few exceptions. It fairly describes the evolution of RAMP over the years and, with the unfortunate exception of the aquatic vegetation and the acid sensitive lakes programs, which were not addressed, it is a good description of what was done. The problems with the report are found in lack of details of methods, failure to describe rationales for program changes, examples of inappropriate statistical analysis, and unsupported conclusions.

Although the Five Year Report was compiled in a satisfactory way, the *content* of the report raised significant concerns with the reviewers about the integrity of the RAMP

Program itself. In the current state, RAMP is not in a position to measure and assess development-related change locally or in a cumulative way. Reviewers reported serious problems related to scientific leadership and a lack of integration and consistency across components with respect to approach, design, implementation, and analysis. Reviewers also reported a lack of an overall regional plan, that clear questions were not been addressed in the monitoring and that there were sometimes significant shortfalls with respect to statistical design of the individual components. Although RAMP appears to recognize that characterization of variability, assessment of regional trends and cumulative effects, and verification of EIA predictions are essential objectives for the program, there is no clear direction on how to achieve and integrate these objectives, despite good existing examples in other national and regional monitoring programs.

There are several levels of recommendations that were provided in this review. Individual component templates and summary reports contain recommendations on details specific to that component. However, after the Design and Integration Team compiled these component-based recommendations, deficiencies, concerns and "theme" areas emerged that were common threads across components. These theme recommendations are provided below and are the most important considerations for RAMP.

Recommendations

The following recommendations are meant to provide a more reliable and systematic approach to aquatic monitoring:

I. Organizational Recommendation on Scientific Leadership

We recommend that RAMP establish a new independent position of project scientific leader reporting to the RAMP Steering Committee and responsible for the overall scientific design of the program and ensuring program quality and relevance through independent peer review. RAMP should also establish an ongoing system of independent scientific input to the program through (1) informal or formal commentary on early ideas and initial plans; (2) workshops and planning sessions that involve independent researchers, RAMP contractor staff and RAMP technical committee members in interchange and debate; (3) formal written review of monitoring plans; and (4) formal review of progress on a periodic basis.

Several findings support the need for a new organizational structure: the need for a clearly delineated overall regional monitoring plan with clear questions to be addressed; the need for establishing a core level of consistency across program components; the need for ongoing independent scientific input into planning programs; the need for ongoing independent scientific peer review of progress (e.g. see the vegetation component); a lack of integration between individual components of the program; and the initiation of program elements that lie outside the capacity/responsibility of the contractor. The RAMP program has been designed by committee consensus and the program has been reactionary and ever-changing. This has resulted in a program where few stations have been sampled consistently over time, consistently across components and using consistent methods. Under these circumstances it will not be possible for the

RAMP to meet its 3 primary objectives. We feel these problems are the result of a lack of a scientific leader.

An independent scientific leader reporting to the Steering Committee would be responsible for the overall scientific design of the program and would work with the main contractor, other minor contractors and outside specialists to lead strategic planning and evaluation. This individual's position would be full-time and responsibilities would be more than a simple liaison officer between the RAMP Steering Committee and the contractor. This individual would have an aquatic, scientific background, hold a strategic vision, and be familiar with EIA approaches and programs such as EEM, RCA, and federal and provincial monitoring. This individual would be the strategic planner of the program and would require adequate resources to do the task. Independent scientific leadership is needed and it should not rest with the lead coordinator for the contractor. The contractor is responsible for delivering the program and reporting on it. The contractor should not be responsible for the overall design or the evaluation of progress, which would create a conflict of interest.

Some of the reviewers suggested an alternative model to the single contractor model, e.g. more along the lines of the NRBS, with a secretariat that provided scientific leadership and coordination and many individual private contractors, and university and government researchers carrying out the projects. We disagree because that model is more suited to individual projects, rather than a long-term, integrated monitoring program.

Several component groups recommended the establishment of an external science advisory panel (e.g. climate and hydrology, fisheries, vegetation), but we recommend against such an option. Given the uncertainty that exists in the management decisions that will be necessary, we feel emphasis should be placed on more flexible, adaptive approaches in which the expertise and knowledge of the wider scientific community can be called upon. Problems with an ongoing advisory board include: (1) board advice is restricted to the expertise of the board members. Expanding the size of the board increases the expertise but smaller boards function better in terms of member participation and overall output; (2) individuals involved in the initial plans cannot be expected to be as objective as those outside the process during reviews of progress; (3) the ongoing time commitments for board members can become too great, with the result that members become unable to commit time and effort at the desired level; and (4) board member ennui after repeated input on the same issues. Issue-specific scientific input may be more difficult to organize than an ongoing advisory board but the results are likely to be more effective when the participation is tailored to the issue. Scientists thrive on novelty and are more ready to participate in specific planning and review exercises on a periodic, rather than ongoing, basis. As well, they are more willing to take part when their time commitments can be clearly defined, their specific expertise is obviously useful, and acceptance of their advice is more probable.

II. Primary Technical Recommendations

1. Adoption of an Ecosystem Approach and Decision-Making Strategy

We recommend that RAMP adopt a strategic, integrated, regional monitoring design and decision-making strategy for measurement of development-related change at an ecosystem level while incorporating site-specific needs. Monitoring must fit within the context of an adaptive management framework and focus beyond project-specific needs. This approach should:

- Consider how decisions on change will be made and the information that is required to make those decisions. For example, what indicators will be measured to assess a particular development activity? What will the indicator be compared against to determine when a change has occurred? Will changes of a certain magnitude and direction trigger a specific line of decisions or an approach to greater monitoring intensity? What will the process be if water quality indicators show a change but no change was measured in fish indicators?
- Consider the development projections to 2020 in the oil sands area and select strategic monitoring locations accordingly. Depending upon the watershed, development level, and physical, chemical, and biological characteristics the monitoring approach can be customized. Sampling intensity and frequency can also be customized;
- Integrate RAMP components (i.e. hydrology, water and sediment quality, benthic invertebrate community structure, fish population health, aquatic vegetation and acid sensitive lakes) at integrated monitoring stations;
- Use adaptive feedback loops within and among components for constant examination of experimental designs and results; changes should be made to the program based on solid results rather than on speculation;
- Show clear links to objectives and have clearly stated hypotheses or testable study objectives; and
- Ensure that all terms, especially statistical ones, are defined and used precisely in reports, and a glossary for all component subject areas be produced as an aid to authors and readers of reports. Precise use of terms aids understanding.

RAMP has changed from year to year. This lack of consistency and strategy has severely limited the ability of RAMP to monitor the environment relative to existing and future development pressures. This comment was common across components including acid sensitive lakes, benthic invertebrates, fisheries, water quality and aquatic vegetation. Development projections to 2020 have been available since the inception of RAMP and extensive information on development has been submitted by independent proponents under the EIA process. The goal of RAMP should be to describe key environmental components, overlay development-related stressors on those environmental components and determine if the change in one can be explained by the other. The monitoring program must be designed to collect environmental information capable of detecting change due to a specific development including selection of appropriate parameters and indicators, and collection at appropriate times and frequencies. For example, investigators stated that “wetland vegetation has been documented as an important biomonitoring parameter for examining potential effects to wetland systems”. Yet they failed to spell out exactly how the vegetation monitoring will enable the investigators to detect effects of oil

sands development. Investigators must state what constitutes unacceptable change in an environmental component in response to oil sands development.

A strategic vision cannot not be implemented unless there is scientific leadership of RAMP as discussed in Recommendation 1. After five years, and considering the development pushing ahead in the oil sands, it is alarming that the main monitoring program for the area significantly lacks strategic direction and scientific process.

2. Adoption of Effects-Based Monitoring within the Strategy

We recommend that RAMP orient its efforts towards effects-based monitoring. The objective should be to document environmental change occurring as a result of development, not to carry out descriptive studies. Included in the effects-based approach should be the following:

- Selection of key response indicators for each RAMP component, based upon potential changes resulting from oil sands development;
- On-going synthesis of information related to development pressures including type of development activity, location of activity, stressors released, effects predicted, assumptions used in predictive tools, location of modeling nodes, etc. A monitoring program designed to monitor development-related change cannot do so in the absence of information on the development. This was recognized as a significant shortfall of the RAMP. Reviewers recognized that much of this information is likely included in the EIA reports. However, effects-based monitoring mandates an on-going comparison between development activities and environmental condition. One without the other will not measure development-related change;
- Establishing a core level of consistency for sample station selection, indicator selection, sampling frequency and timing that does not change from year to year;
- Selection of reference and "low-impact" stations within or outside the Region for each component subject area. Those subject areas that can go into an established biomonitoring program (see below) will get this benefit automatically;
- Use of biostatistical analyses that report statistical confidence levels and power analyses for indicators of change. These statistical results are critical to assist with interpretation of the environmental changes to establish confidence in the decision-making strategy;
- Consideration of the knowledge and understanding gained from other successful effects-based monitoring programs that measure development-related change relative to natural variability; for pertinent subject areas such as water quality, benthos, fish and possibly aquatic vegetation, a bona fide, regional biomonitoring program (Environmental Effects Monitoring [EEM] or the Reference Condition Approach [RCA]) should be initiated; and
- Incorporation of other existing regional information such as NRBS, NREI, PERD, EEM, the Muskeg River design initiative (CEMA) and information collected independently by industry. Future periodic summary reports, such as the next Five Year Report, should incorporate monitoring results and studies from programs other than RAMP, if the information contributes to the objectives.

3. Testing Environmental Impact Assessment (EIA) Predictions
We recommend that RAMP complete an exercise to test predictions from already completed EIAs using actual data generated on a site or sites. As a first step in this evaluation, RAMP should prepare a synthesis or summary, on a project-specific basis, of what the impact predictions were for different project activities, including location and timing of impact and Valued Ecosystem Components (VECs) affected.

Conducting a follow-up by verifying impact predictions using real data would be a valuable exercise to illustrate exactly what the deficiencies and gaps are in the existing monitoring program and what needs to be done so that predictions can be verified. The Five Year Report did not attempt to verify EIA predictions.

4. Development of an Information Management System
We recommend that RAMP establish a comprehensive information management and assessment system, including an electronic database management system that would enable electronic reporting of raw data in a standard and consistent format, interchange of data among component subject areas, and on-going assessment of data using consistent analyses.

The current method of reporting and data integration is not sustainable, and access to information by RAMP users cannot be facilitated using this approach. Reviewers found table after table of data too difficult to synthesize, and the value of the data was lessened by this reporting structure. This recommendation, however, does not pertain to simply a database with query capabilities. RAMP requires a spatially explicit (GIS-based) system where development layers can be overlain with environmental information for all components and stations. There is a requirement for the data to be graphed using standard formats over time and space, and for the data to be exportable for statistical analyses. There are several on-going initiatives within the region that RAMP could benefit from including the federal EcoAtlas-CE system and the provincial information management initiatives. RAMP information should not be placed into a system that operates independently of these other systems. RAMP depends heavily upon federal and provincial monitoring data (e.g. water quality program) and should make efforts to integrate any system they develop. RAMP should also incorporate other industry data that are being collected independently of the current RAMP program. Participation in an existing information management system will ensure cost-effectiveness and continuity in data management and access among contractors.

5. Increased Emphasis on the Athabasca River as a Priority Watershed
We recommend that RAMP use the Athabasca River as a central focus for monitoring across component subject areas because it is the largest and most important aquatic ecosystem in the region and the natural recipient of the effects of oil sands development.

There is currently no ability within RAMP to assess oil sands development impacts on the Athabasca River in an integrated way. Hydrology data on the Athabasca River were described by reviewers as being significantly limited. Water quality monitoring was conducted at sites too far separated and with inadequate statistical replication to measure

changes due to oil sands development independent of the river continuum (natural changes). Benthic invertebrate monitoring was conducted in the early 1990s but was discontinued due to sampling challenges. Fish work was conducted but there is no integration of this component with the other RAMP components. Given that other monitoring programs have operated successfully on the Athabasca, and the river is a critical integrator of potential impacts, this is an unexplained gap. Development of the strategic plan and effects-based monitoring design should be a first priority for the Athabasca River.

III. Secondary Technical Recommendations

1. Contributions to New Knowledge

We recommend that RAMP recognize the importance of creating new knowledge and incorporating this knowledge into the monitoring program through an adaptive management framework.

The primary purpose of RAMP is to produce knowledge of how the ecosystem is changing over space or time and/or in response to impacts. A side benefit to monitoring can be the production of new functional knowledge or understanding, which will only result when the data produced by monitoring are used to test an explicit hypothesis. If monitoring is to contribute to the long-term assessment of aquatic resources then it must take place as part of a specific experimental design. Reviewers felt that there is an unrealized opportunity that is not being met for creation of new scientific understanding from RAMP monitoring. Comments about RAMP contributions to new knowledge can be found in the climate and hydrology and benthic invertebrate reports. RAMP could be producing results that contribute to regional, national or international understandings of spatial and temporal trends and cumulative effects and about the nature of impacts on ecosystem function. In so doing it could contribute to better models and better prediction of environmental impacts in the future but, as currently operated, it will not do so, until a better-designed, overall strategic monitoring framework is in place.

2. Traditional Ecological Knowledge (TEK)

We recommend that RAMP actively promote the use of TEK by incorporating it into the design of scientific programs. Key indicators for future monitoring and the interpretation of results need to be identified, and specific, ongoing programs should be devoted to observing changes in these key indicators.

We considered that, even though five of the eight RAMP objectives (Appendix I) were not the focus of the Five Year Report, there should be some evidence in the content of the programs that would tell us whether those objectives were being addressed at all. We asked the reviewers for comments on those objectives as they related to the discipline they were reviewing. Comments were most often received on TEK. Several of the reviewers felt that TEK could be contributing to the program. However, there is no evidence anywhere that it has been considered other than in some of the statements on objectives early in program development. It is assumed that some of the parameters measured in the water-quality, vegetation or fisheries components were identified by

stakeholders as VECs in the environmental assessment process. However, the report does not include any information on which parameters were included. Thus, it is not clear if TEK was used as a basis for parameter selection. A separate review of a fish abnormalities study was completed and, although the original concerns came from local residents, there was no evidence that their knowledge had been used in any way. For fisheries programs in particular, local knowledge can provide information on what species have been historically important, and during what seasons they are present, and it can contribute to the overall understanding of the functioning of the system (e.g. are these species migratory with harvests from outside the immediate system, and are they locally important as well?).

A recent government report on science advice for government effectiveness (CSTA, 1999) states that decision-makers should be taking due weight of the traditional knowledge of local peoples. It goes on to say that traditional knowledge, like scientific knowledge, needs to be subjected to due diligence, including rigorous internal and external review and assessment. It is clear to us that RAMP has not taken account of traditional knowledge to the extent one might expect for a study of this nature, especially since it is one of the stated program objectives. Incorporation of TEK with western science needs to be addressed in the ecosystem approach and decision-making strategy.

3. Publications

We recommend that RAMP initiate a policy of encouraging individuals and the contractor to publish monitoring data and new knowledge in established technical and primary publications as well as in-house reports. RAMP should also establish a RAMP Technical Report Series for wider distribution of monitoring results within the region, provincially and nationally.

Comments about the potential usefulness of RAMP primary and technical publications were made in the water-quality, benthic invertebrates, vegetation and acid sensitive lakes reports. We strongly believe that the results of the RAMP program should be widely disseminated in a more formal manner. High publication standards require high monitoring standards. Publication of results imposes more scientific rigor on the monitoring program, it adds to credibility of the program, it increases exposure of project managers to current scientific information in other areas, and it contributes new information to the program itself. It also adds to the personal capacity and credibility of the individuals involved in the monitoring, resulting in employees who are more satisfied in their jobs.

The proposed technical report series should be structured like some of the government data or technical report series (e.g. Canadian Manuscript Report of Fisheries and Aquatic Sciences). It would be a series of reports generated from the information management system on an on-going basis. This effort would not be onerous if designed properly.

There is a formal procedure for establishing a new series of reports. An ISSN should be included in each report. Numbers can be applied for online at http://www.nlc-bnc.ca/issn/index-e.html. An electronic copy should be sent to observe the legal

requirement for filing a depository copy with the National Library of Canada (see http://www.nlc-bnc.ca/6/25/index-e.html.) and copies should be sent to regional, provincial and federal libraries to ensure cataloguing in environmental databases.

Conclusion

The above are general recommendations that we feel need to be implemented for RAMP. Other general and specific component subject recommendations are presented in the individual narratives above and in the template reports in Appendix IV. There are a number of individual recommendations that could be implemented immediately. We recognize that RAMP is entering initial planning for 2005, so there will be a temptation for RAMP Steering Committee members, RAMP Technical Committee members and the contractors to seize upon "favored" recommendations for immediate action.

We would urge caution in this respect. We have tried to emphasize that there are some overall structural changes that need to take place within the program. The primary need is for scientific leadership and input to a strategic planning process that treats the program as a single entity not as a series of individual components. To begin immediate implementation of minor specific changes risks continuation of a pattern that has created some of the problems with RAMP in the first place, i.e. lack of continuity and change of programs without sound justification.

We have not identified specific research recommendations because of our belief that the core monitoring program needs to be changed in a major way (see above), and should be the focus of intense effort over the short term. Thus, specific research recommendations should follow reorganization of the monitoring program.

ACKNOWLEDGEMENTS

Most importantly we would like to thank and acknowledge the work of the individual component reviewers: Neil Arnason, Jan Barica, Brian Brownlee, Uwe Borgmann, Martin Carver, Nancy Glozier, Kelly Munkittrick, Carl Schwarz, Brian Souter, Stephanie Sylvestre, Alan Thomson, Michael Turner, John Post, and Marley Waiser. The reviewers were all experienced with reviewing scientific reports for journal publication and familiar with reviewing research proposals for agencies such as NSERC but large-scale independent scientific peer review of aquatic environmental monitoring programs is not a common occurrence. This task was larger and more complex than any of us had imagined and we appreciate their hard work, perseverance and understanding as we strove to produce a understandable and useful assessment of this important program.

We would also like to thank the RAMP Review Team -- Bryan Kemper, Christine Brown, Preston McEachern, and Mark Spafford -- for their initial direction and support during the process and the RAMP Technical Sub-committee for their input and suggestions following the October 23, 2003 progress report.

We want to acknowledge the work of the contractor, Golder Associates Ltd., of Calgary who were responsible for carrying out the RAMP until 2002 and responsible for preparing the annual reports and the Five Year Report. The Golder staff are to be commended for their efforts in operating this very important program, responding to the problems and issues in the environment of the Oil Sands Region and the complex demands of the regulators, the oil companies and the other stakeholders that make up the RAMP Steering Committee.

We would like to thank and acknowledge the work of Donna Laroque of Winnipeg. She was responsible for formatting the template reports, and inputting, and formatting much of the overall report.

We would like to acknowledge the following individuals who provided advice and/or assistance to components of the review: Dr. Michael Church, University of British Columbia; Mr. Tim Davis, Water Survey of Canada; Dr. Wayne C. Huber, Oregon State University, Dr. Robert Newbury, Newbury Hydraulics, D. W. Schindler, University of Alberta; Bill Hume, Environment Canada; and Lawrence Cheng and Preston McEachern, Alberta Environment, Dave Ballagh, Saskatchewan Environment, Dr. Art Tautrz, BC Fisheries Branch.

BIBLIOGRAPHY

Reports Reviewed

RAMP (Oil Sands Regional Aquatics Monitoring Program). 1997-2001. Annual Reports 1997-2001. Submitted to RAMP Steering Committee by Golder Associates Ltd., Calgary, AB.[3]

RAMP (Oil Sands Regional Aquatics Monitoring Program). 2000b. Fish abnormalities. Submitted to RAMP Steering Committee by Golder Associates Ltd., Calgary, AB. 39 p.

RAMP (Oil Sands Regional Aquatics Monitoring Program). 2002a. Annual Report 2002. Submitted to RAMP Steering Committee by Golder Associates Ltd., Calgary, AB.[4]

RAMP (Oil Sands Regional Aquatics Monitoring Program). 2002b. Program design and rationale: version 2. Submitted to RAMP Steering Committee by Golder Associates Ltd., Calgary,AB. 298 p.

RAMP (Oil Sands Regional Aquatics Monitoring Program). 2003. Five Year Report. Submitted to RAMP Steering Committee by Golder Associates Ltd. Calgary,AB. 602 p.

Reports Cited in Text

Bailey, R.C., R.H. Norris, and T.B. Reynoldson. 2004. Bioassessment of freshwater ecosystems using the Reference Condition Approach. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Bisbal, G.A. 2002. The best available science for the management of anadromous salmonids in the Columbia River basin. Canadian Journal of Fisheries and Aquatic Sciences 59:1952-1959.

Borgmann, U., W.P Norwood, T.B. Reynoldson and F. Rosa. 2001. Identifying cause in sediment assessments: bioavailability and the sediment quality triad. Canadian Journal of Fisheries and Aquatic Sciences 58:950-960.

Borgmann, U. 2003a. Derivation of cause-effect based sediment quality guidelines. Canadian Journal of Fisheries and Aquatic Sciences 60:352-360.

Borgmann, U. 2003b. Assessing metal impacts in sediments: key questions and how to answer them. In M. Munawar (editor) Sediment quality assessment and management: insight and progress. Ecovision World Monograph Series, Aquatic Ecosystem Health and Management Society, pp. 23-28.

---

[3] Reports are available on CD from Golder Associates Ltd., 1000-940 6th Avenue S.W., Calgary, Alberta, Canada, T2P 3T1, Att. Kym Fawcett e-mail kfawcett@golder.com

[4] RAMP (Oil Sands Regional Aquatics Monitoring Program). 2002. Annual Report 2002. Submitted to RAMP Steering Committee by Golder Associates Ltd., Calgary, was examined by the acid sensitive lakes reviewers.

Crosley, R.W. 1996. Environmental contaminants in bottom sediments, Peace and Athabasca River Basins, October, 1994 and May, 1995. Project Report No. 106. Northern River Basins Study, Edmonton, AB.

CSTA (Council of Science and Technology Advisors). 1999. Science Advice for Government Effectiveness (SAGE). Available from Science and Technology Directorate, Industry Canada, Ottawa, ON. 11 p.

Dorcey, A.H.J. and K.J. Hall. 1981. Setting ecological research priorities for management: The art of the impossible in the Fraser estuary. West Water Research Centre, the University of British Columbia, Vancouver, BC. 78 p.

Dubé, M.G. 2003. Cumulative effects assessment in Canada: a regional framework for aquatic ecosystems. Environmental Impact Assessment Review (in press).

Environment Canada. 1997. Benthic invertebrate community expert working group final report: recommendations from cycle 1 review. EEM/1997/7. National EEM Office, Science Policy and Environmental Quality Branch, Environment Canada, Ottawa, ON.

Environment Canada. 1998. Pulp and paper technical guidance for aquatic environmental effects monitoring. EEM/1998/1. National EEM Office, Science Policy and Environmental Quality Branch, Environment Canada, Ottawa, ON.

Environment Canada. 2001. Metal mining guidance document for aquatic environmental effects monitoring. National EEM Office, Science Policy and Environmental Quality Branch, Environment Canada, Ottawa, ON. (Available from: www.ec.gc.ca/eem)

Fleishman, E. 2001. Moving scientific review beyond academia. Conservation Biology 15:547-549.

Glozier, N.E., J.M. Culp, T.B. Reynoldson, R.C. Bailey, R.B. Lowell, and L. Trudel. 2002. Assessing metal mine effects using benthic invertebrates for Canada's environmental effects program. Water Quality Research Journal of Canada 37:251-278.

Helgen, J.C. and M.C. Gernes. 2001. Monitoring the condition of wetlands: Indexes of biological integrity using invertebrates and vegetation. In R.B. Rader, D.P. Batzer and S.A. Wissinger (editors). Bioassessment and management of North American freshwater wetlands. John Wiley and Sons, New York. pp.167-186.

Karr, J.R. and E.W. Chu. 1999. Biological monitoring and assessment: Using multimetric indexes effectively. Island Press, Cobelo, CA.

Meffe, G.K., P. Dee Boersma, D.D. Murphy, B.R. Noon, H.R. Pulliam, M.E. Soulé, D. Waller. 1998. Independent scientific review in natural resource management. Conservation Biology 12:268-270.

Peterman, R.M. 1990. Statistical power analysis can improve fisheries research and management. Canadian Journal of Fisheries and Aquatic Sciences 47:2-15.

Reid, L.M. 1993. Research and cumulative watershed effects. General Technical Report PSW-GTR-141. USDA Forest Service Pacific Southwest Research Station. 118 p.

Rader, R.B. and D.K. Shiozawa. 2001. General principles of establishing a bioassessment program. In R.B. Rader, D.P. Batzer and S.A. Wissinger (editors). Bioassessment and management of North American freshwater wetlands. John Wiley and Sons, New York, pp.13-44.

Reynoldson, T.B., R.C. Bailey, K.E. Day, and R.H. Norris. 1995. Biological guidelines for freshwater sediment based on Benthic Assessment of SedimenT (the BEAST) using a multivariate approach for predicting biological state. Australian Journal of Ecology 20:198-219.

Richardson, C.J. and J. Vymazal. 2001. Sampling macrophytes in wetlands. In R.B. Rader, D.P. Batzer and S.A. Wissinger (editors). Bioassessment and management of North American freshwater wetlands. John Wiley and Sons, New York, pp.297-338.

Rosenberg, D.M. and V.H. Resh. 1993. Introduction to freshwater biomonitoring and benthic macroinvertebrates. In D.M. Rosenberg and V.H. Resh (editors). Freshwater biomonitoring and benthic macroinvertebrates. Chapman and Hall, New York, pp.1-9.

Rosenberg, D.M., T.B. Reynoldson, and V.H. Resh. 1999. Establishing reference conditions for benthic macroinvertebrate monitoring in the Fraser River catchment, British Columbia, Canada. DOE-FRAP 1998-32. Fraser River Action Plan, Environment Canada, Vancouver, BC.

US Environmental Protection Agency (USEPA). 2002. Methods for evaluating wetland condition: #10 – Using vegetation to assess environmental conditions in wetlands. #6 Developing metrics and indexes of biological integrity. #4 Study Design for Monitoring Wetlands. (On EPA website: http://www.epa.gov/waterscience/criteria/wetlands/)

Vecchia, A.V. 2003. Water-quality trend analysis and sampling design for streams in North Dakota, 1971-2000. Water-Resources Investigations Report 03-4094. U.S. Department of the Interior, U.S. Geological Survey. 73 pp. (Available from www.usgs.gov)

Walker, S.L., K. Hedley, and E. Porter. 2002. Pulp and paper environmental effects monitoring in Canada: an overview. Water Quality Research Journal of Canada 37:7-19.

Wright, J.F. 1995. Development and use of a system for predicting the macroinvertebrate fauna in flowing waters. Australian Journal of Ecology 20:181-197.

Wright, J.F., D.W. Sutcliffe, and M.T. Furse (editors). 2000. Assessing the biological quality of fresh waters. Freshwater Biological Association, Ambleside, Cumbria, UK.

WSOCD (Washington State Office of Community Development). 2002. Citations of recommended sources of best available science for designating and protecting critical areas. 89 p. Olympia, Washington. (Available from: Washington State Office of Community Development, P.O. Box 48350, Olympia, WA 98504-8350.)

APPENDIX I: OBJECTIVES OF THE OIL SANDS REGIONAL AQUATIC MONITORING PROGRAM (RAMP)

These objectives are taken from the Terms of Reference of RAMP and from the Five Year Report. The focus of the Five Year Report and this review is on the first three objectives. The specialist reviewers were also asked to note whether, based on their reading, the program also addressed the last five objectives.

1. **Characterizing Existing Variability** - To collect scientifically defensible baseline and historical data to characterize variability in the oil sands area. (Note from Design and Integration Team - The capacity to detect change was of particular importance for reviewers to consider.)
2. **Detecting Regional Trends and Cumulative Effects** - To monitor aquatic environments in the oil sands area to detect and assess cumulative effects and regional trends. (Note from Design and Integration Team - The capacity to detect cumulative effects and trends in consideration of new disturbances was of particular importance for reviewers to consider.)
3. **Monitoring to Verify EIA Predictions** - To collect data against which predictions contained in environmental impact assessments (EIAs) can be verified.
4. **Monitoring to Meet Regulations** - To collect data that may be used to satisfy the monitoring required by regulatory approvals of developments in the oil sands area.
5. **Traditional Ecological Knowledge** - To recognize and incorporate traditional knowledge (including Traditional Ecological Knowledge and Traditional Land Use Studies) into the monitoring and assessment activities.
6. **Communication** - To communicate monitoring and assessment activities, results and recommendations to communities in the Regional Municipality of Wood Buffalo, regulatory agencies, environmental committees/organizations and other interested parties.
7. **Flexibility and Adaptability** - To design and conduct various RAMP activities such that they have the flexibility to be adjusted, on review, to reflect monitoring results, technological advance and community concerns.
8. **Cooperation** - To seek cooperation with other relevant research and monitoring programs where practical, and generate interpretable results which can build on their findings and on those of historical programs.

APPENDIX II: BIOGRAPHIES OF INDIVIDUAL REVIEWERS OF THE RAMP PROGRAM

Neil Arnason, Ph.D
Department of Computer Sciences
University of Manitoba
Winnipeg, MB
phone: 204-474-6918
e-mail: arnason@cs.umanitoba.ca

Dr. Neil Arnason received his B.Sc. and M.Sc. in statistics at the University of Waterloo and a Ph.D. in biometrics from Edinburgh (1971). Since then, he has taught courses in statistics, quantitative ecology, and computer science at the University of Manitoba where he is now a Full Professor in the Department of Computer Science. His research is in population estimation methods, and he has published papers (in Biometrics, Journal of Wildlife Management, Canadian Journal of Fisheries and Aquatic Sciences, among other journals) and software (from his website www.cs.umanitoba.ca/~popan) related to the estimation of animal abundance and migration rates. He consults with biologists from all over the world on population survey design and data analysis. He is a member of the Statistical Society of Canada and was president of the Biostatistics Section (1998-1999) and currently is active on the Professional Accreditation committee. He is also a member of the Biometrics Society, the American Statistical Society and a Fellow of the Royal Statistical Society.

Burton Ayles, Ph.D.
B. Ayles and Associates
255 Egerton Road
Winnipeg, MB R2M 2X3
phone: 204-257-4453
fax: 204-257-4453
e-mail: aylesb@escape.ca

Dr. Burton Ayles received his B.Sc. and M.Sc. in zoology and genetics from the University of British Columbia and his Ph.D. in fisheries genetics from the University of Toronto (1972). He worked for 25 years for the federal Department of Fisheries and Oceans as a research scientist and manager at all levels in the organization, including, Regional Planner, Regional Director of Operations, Regional Director of Research, and Regional Director General for the Central and Arctic Region (Prairies, Ontario and the Arctic). As a senior DFO manager he was intimately involved in design and evaluation of regional and national science and technology programs on an ongoing basis. In 1998 Dr. Ayles took an early retirement from the federal government and established B. Ayles and Associates Fisheries and Environmental Consulting, providing advice and evaluation on policy, research and management. Dr. Ayles is a Canada Member of the Canada/Inuvialuit Fisheries Joint Management Committee which, with the Department of Fisheries and Oceans and the Inuvialuit Game Council, is responsible for management of fisheries and marine mammals in the western Canadian Arctic. He has been active planning and organizing workshops, planning sessions and reviews for a range of fisheries and aquatic environmental activities including: Arctic fisheries and oceans

research, water quality in the prairies, and sustainability of the Muskeg River, amongst others.

Jan Barica, Ph.D, D.Sc.
2153 Lincoln Court
Burlington, ON
L7P 3S4
phone: 905-335 1633
e-mail: jbarica@cogeco.ca

Dr. Jan Barica's career with the government of Canada's Department of Fisheries and Oceans and Department of Environment, over 30 years as a research scientist and research manager, focused on management and restoration/rehabilitation of lakes, reservoirs and river basins. In Canada he worked on hypereutrophic lakes in western Canada and their utilization for fish culture; algae control, manipulation of algal blooms by nutrient-ratio adjustment, aeration and dredging; and on basin-wide eutrophication controls in the Great Lakes and their tributaries, lake-wide management programmes for Lake Ontario and Erie, remediation action plans in the Areas of Concern (Hamilton Harbour, Severn Sound); water quality surveillance programmes; and long-term data interpretation. Throughout his career he was actively involved in many international activities ranging from Iraq to the Philippines, Thailand, South America and many countries in eastern Europe. Since his retirement in 1999 he has been active in UNEP-UNDP-GEF-IDRC programs in Ukraine, Belarus and Russia on water quality, biodiversity and strategic planning.

Brian Brownlee, Ph.D.
Environment Canada
Room 200, 4999-98 Ave
Edmonton, AB T6B 2X3
phone: 780-951-8745
e-mail: brian.brownlee@ec.gc.ca

Dr. Brian Brownlee received his B.Sc. and M.Sc. in chemistry from the University of Alberta and his Ph.D. in synthetic organic chemistry from the University of New Brunswick (1971). He is a research scientist with Environment Canada's National Water Research Institute and has over 30 years experience in research on water quality related issues in Canada. Specific areas studied include oil sands contaminants, taste and odour in drinking water supplies, urban runoff, benzothiazoles, nurient dynamics, and pulp mill contaminants. His research has covered an extensive geographic range including the Great Lakes, small prairie lakes, the Alberta oil sands, northern Alberta rivers and southern Ontario lakes and streams, amongst others.

Uwe Borgmann
National Water Research Institute
Environment Canada
867 Lakeshore Road, P.O. Box 5050
Burlington, Ontario, Canada, L7R 4A6
phone: 905-336-6280

fax: 905-336-6430
email: uwe.borgmann@ec.gc.ca

Dr. Borgmann received his M.Sc. in zoology and oceanography from the University of British Columbia and his Ph.D. in biology from the University of Ottawa (1975). He was a research scientist with the Department of Fisheries and Oceans for 20 years and since 1996 he has been a research scientist with Environment Canada's, National Water Research Institute. His current research interests include invertebrate toxicology with emphasis on metals; relationship between metal bioaccumulation and toxicity; and application of bioaccumulation and other bioavailability measures to environmental risk assessments of metals. He is also an adjunct professor at the University of Waterloo and has supervised several M.Sc. and Ph.D. students. He is active on several scholarly and professional societies including currently acting as Associate Editor, Canadian Journal of Fisheries and Aquatic Sciences, and as a member of the editorial board of the journal, Aquatic Ecosystem Health and Management.

Martin Carver, Ph.D., P.Eng., P.Ag.
Carver Consulting
#1 - 4925 Marello Road
Nelson, BC. V1L 6X4
phone: 250 352-1187
fax: 250 352-1197
e-mail: carver@netidea.com

Dr. Martin Carver received his M.A.Sc. from the University of Waterloo, and his Ph.D. in 1997. He is an international consultant with 13 years' experience in water resources and land management emphasizing watershed condition, water quality, forestry and agriculture. His expertise includes: hydrologic/fluvial geomorphological research and modeling; development of forest hydrology and watershed management assessment procedures; geomorphological and hydrological field measurements and monitoring; riparian and hydrologic assessments of streams; water quality assessment and diagnosis and watershed restoration. Recent activities include preparation and delivery of a three-day course in Equador on the management of tropical mountain watersheds; review of technical studies for Connor Creek watershed to recommend priority hydrologic mitigation/restoration activities; preparation of an integrated riparian management plan for Arrow Creek – a large high-value watershed in Creston, BC; conducted and reviewed watershed assessments for watersheds in the Nelson Forest Region and in Ecuador.

Dubé, Monique, Ph.D
National Water Research Institute, Environment Canada
11 Innovation Blvd.
Saskatoon, SK S7N 3H5
phone: 306-975-6012
fax: 306-975-5143
e-mail: monique.dube@ec.gc.ca

Dr. Monique Dubé is a Research Scientist in the Cumulative Impacts on Aquatic Ecosystems Group at the National Water Research Institute of Environment Canada in Saskatoon, SK. She is also an Adjunct Professor at the Toxicology Centre at the

University of Saskatchewan and a member of the Canadian Rivers Institute. Her expertise includes effects assessment of industrial and municipal effluents on riverine food webs and development of mesocosm and stable isotope approaches for environmental effects monitoring. Recent activities include: development of a regional cumulative effects assessment framework for aquatic ecosystems and an associated software system for framework implementation; membership in the National Environmental Effects Monitoring Science Committee on pulp and paper and metal mining industries across Canada; and participation as an invited outside expert in a workshop on the sustainability of the Muskeg River ecosystem.

Ms. Nancy E. Glozier, M.Sc.
Aquatic Ecosystems Scientist
Environment Canada, PNR Wildlife Research Centre
115 Perimeter Road, Saskatoon, SASK S7N 0X4
phone: 306- 975-6057
e-mail: nancy.glozier@ec.gc.ca

Nancy Glozier received both her B.Sc., in zoology, and her M.Sc. (1989), in aquatic ecology, from the University of Calgary. She joined Environment Canada as a research support technician and since 2002 she has been aquatic ecosystems research scientist. Her recent research interests include environmental fate and effects of pulp and paper mill effluents, use of mesocosms for assessment of cumulative effects in algal, benthic and fisheries communities, and applications of the use of artificial streams for Environmental Effects Monitoring (EEM) amongst others. She is currently a member of Environment Canada's National Science committee and a member of the EEM benthic expert subgroup for metal mining.

Kelly Munkittrick, Ph.D.
Department of Biology, University of New Brunswick
P.O. Box 5050, Tucker Park Road
Saint John, NB E2L 4L5
phone: 506-648-5825
fax: 506-648-5811
e-mail: krm@unb.ca

Dr. Kelly Munkittrick received his Ph.D. in Toxicology from the University of Waterloo, and spent time in private industry and consulting environments before joining the Federal Government in 1990. He currently holds a Tier 1 Canada Research Chair in Ecosystem Health Assessment at the Department of Biology, University of New Brunswick, Saint John. Prior to his appointment, he worked for 10 years for the federal government as a Project Chief with the Ecosystem Health Assessment Project at Environment Canada's National Water Research Institute, and as a Research Scientist with Fisheries and Oceans' Great Lake Laboratory for Fisheries and Aquatic Sciences. His research interests are on environmental health assessment, cumulative effects assessment and the impacts of industrial discharges on wild fish populations. He has co-chaired interdisciplinary working groups related to Environmental Effects Monitoring, and is a past co-chair of both the Environment Canada and the Canadian (5NR) Interdepartmental Endocrine Disruptor Working Group.

Dr. John Post, Ph.D.
Department of Biological Sciences, University of Calgary
2500 University Drive
Calgary T2N 1N4
phone:403 220-6937
e-mail: jrpost@ucalgary.ca.

Dr. John Post is an Associate Professor in the Biology Department of the University of Calgary. He received his Ph.D. from York University in Toronto. His research interests include energy dynamics and bioenergetics models, recruitment variability in fishes and population dynamics, dynamics of foraging, growth, spatial behaviour and survival in structured fish populations, the interface between fish biology and fisheries management, and aquatic food web dynamics. His teaching responsibilities have included ecology and evolution, aquatic communities and ecosystems, quantitative biology and ecology of fish. Recent specific activities include: energy allocation strategy in age-0 fish; density dependent inter-cohort interactions and recruitment dynamics; models and a bull trout time series; and recruitment dynamics and size structure in experimental fish populations.

David Rosenberg, Ph.D.
Freshwater Institute
501 University Crescent
Winnipeg, MB R3T 2N6
phone: 204-983-5253
e-mail: rosenbergd@dfo-mpo.gc.ca

Dr. David. Rosenberg received his B.Sc. and Ph.D. from the University of Alberta (1973). He has spent all of his professional life at the Freshwater Institute in Winnipeg, most of it using benthos to monitor environmental disturbance. He has participated in ecological studies of proposed Mackenzie Valley pipelines, diversion of the Churchill River into the Nelson River in northern Manitoba, and experimental reservoir creation at the Experimental Lakes Area in Ontario. More recently, he helped establish a biomonitoring program for the Fraser River in BC, and was part of a group that tried to convince the Federal Government that Canada needs a national biomonitoring program. He retired from the Department of Fisheries and Oceans in September 2001. Recent activities include: Managing Editor of the Journal of the North American Benthological Society; participation as an invited outside expert in a workshop on the sustainability of the Muskeg River ecosystem; and contributing to a planned book on North American river ecosystems

Dr. Carl James Schwarz, Ph.D.
Department of Statistics and Actuarial Science
Simon Fraser University
8888 University Drive
Burnaby, BC V5A 1S6
phone: 604-291-3376
fax: 604-291-4368
e-mail: cschwarz@stat.sfu.ca

Dr. Carl Schwarz is a Professor in the Department of Statistics and Actuarial Science at Simon Fraser University. His research interests are in the use of statistics in ecology - particularly in estimating animal abundances and related parameters using capture-recapture methods and in the design and analysis of environmental monitoring studies.

Brian W. Souter, M.Sc.
Department of Fisheries and Oceans
Central and Arctic Region
Winnipeg MB, R3T 2N6
phone: 204-983-5125
fax: 204-984-2404
e-mail: souterb@dfo-mpo.gc.ca

Brian Souter received his B.Sc. and M.Sc. in microbiology and fisheries from the University of Guelph (1974) and he is a fish health specialist with the Department of Fisheries and Oceans. He has directed the federal fish health certification program in the region since 1977 and he has been the DFO technical representative on the Great Lakes fish health committee of the Great Lakes Fishery Commission since 1980. He works with the National Registry of Aquatic Animal Health to revise the Fish Health Protection Regulations and the Manual of Compliance, and to develop components of the National Aquatic Animal Health Program. He is also the DFO representative on a task force with the mandate to counteract the threat of whirling disease to Alberta's wild and cultured salmon stocks. He has also authored or co-authored several publications on fish health in central Canada.

Stephanie Sylvestre, M.Sc.
Environmental Studies Scientist
Aquatic and Atmospheric Sciences Division
Environment Canada, Environmental Conservation Branch
#201 - 401 Burrard Street,
Vancouver BC, V6C 3S5
phone: 604-664-4099
fax: 604-664-9126
e-mail: stephanie.sylvestre@ec.gc.ca

Stephanie Sylvestre received her B.Sc. from University of Windsor and her M.Sc. from the University of Western Ontario. She joined Environment Canada as an Environmental Studies Officer in 1994 and is now an Environmental Studies Scientist with the Aquatic and Atmospheric Sciences Division in Vancouver. Recent activities include: stream assessments in the Georgia Basin using the reference condition approach for benthic invertebrate monitoring; water quality assessment of agricultural and residential runoff; expanding the use of the benthic invertebrate monitoring approach developed for the Fraser River basin to assess streams in the Georgia Strait basin; PAHs and other contaminants in suspended sediment and water in the Fraser River basin.

Alan R. Thomson,MRM P.Eng
Mountain Station Consultants, Inc.
906 Ninth Street

Nelson, BC V1L 3C3
phone fax 250-352-0016
e-mail: alant@alumni.sfu.ca

Alan R. Thomson, is a principal of Mountain Station Consultants of Nelson, BC, and specializes in resolving natural resource management issues that involve the interaction of aquatic and biological resources. In his 11-year consulting practice, Alan has completed numerous contracts that involve watershed and stream assessments, river hydrology and engineering, design of new and restoration of existing aquatic biota habitats, river channel and bank stabilization, bioengineering, fish migration assessment and passage creation, policy and investigative research, and strategic planning and water quality enhancement and recovery. Recent contracts include: being an expert witness and advisor to the Department of Fisheries and Oceans at environmental impact assessment joint panel hearings concerning oils sands development in northern Alberta, and restoring fish habitats in several streams in British Columbia.

Michael A. Turner, Ph.D.
Fisheries and Oceans Canada
501 University Crescent,
Winnipeg, Manitoba R3T 2N6
phone: 204-983-5215
fax: 204-984-2404
e-mail: turnermi@dfo-mpo.gc.ca

Dr. Michael Turner received his M.Sc. and Ph.D. from the University of Manitoba. He is a Research Scientist of the Department of Fisheries and Oceans in Winnipeg. His primary research is at the Experimental Lakes Area in northwestern Ontario. He has a long history of work on the impact of acidification on lakes. His current limnological research focuses on the littoral ecology of boreal lakes impacted by habitat disruption and by climate variability and change. He also leads a research team studying the recovery of boreal lakes from acidification.

Marley Waiser, Ph.D.
Aquatic Ecosystem Impacts Research Branch
National Water Research Institute
Environment Canada
11 Innovation Boulevard
Saskatoon, SK S7N 3H5
phone: 306-975-5762
fax: 306-975-5143
e-mail: Marley.waiser@ec.gc.ca

Dr. Marley Waiser received her Ph.D. from Napier University in Edinburgh, Scotland. She is a Research Scientist with the Aquatic Ecosystem Protection Research Branch of Environment Canada's National Water Research Institute in Saskatoon. She is also an adjunct professor with the Department of Applied Microbiology, University of Saskatchewan in Saskatoon. Dr. Waiser's research has focused mainly on the microbial ecology and biogeochemistry of prairie aquatic ecosystems including saline lakes and wetlands. Her research has been published in Limnology and Oceanography, Canadian

Journal of Fisheries and Aquatic Research, Archiv für Hydrobiologie, Biogeochemistry and Aquatic Microbial Ecology. Currently, she is investigating the effects of sulfonylurea herbicides on the microbial ecology of prairie wetlands as part of a larger collaborative effort looking at the fate and effects of this new generation of herbicides. Dr. Waiser is also part of a team of scientists who are investigating the relationships between terrestrial and aquatic dissolved organic carbon, with special reference to prairie ecosystems.

Appendix III
Oil Sands Regional Aquatic Monitoring Program (RAMP)
Scientific Peer Review of the Five Year Report (1997-2001): Reviews of Biostatistics

Submitted to:
RAMP Steering Committee

Report Prepared by:
Carl Schwarz
Department of Statistics and Actuarial Science
Simon Fraser University
8888 University Drive
Burnaby, BC, V5A 1S6

cschwarz@stat.sfu.ca
http://www.stat.sfu.ca/~cschwarz

1. INTRODUCTION

The Oil Sands Regional Aquatics Monitoring Program (RAMP) in the Oil Sands Region of north-eastern Alberta was designed to measure baseline environmental conditions and predict effects from proposed developments. RAMP was designed as a long-term monitoring program that incorporates both traditional and scientific knowledge. Specific programs in RAMP were established each year by committees and subcommittees after consultation with industrial, aboriginal, environmental and regulatory stakeholders and expert independent consultants. As the Oil Sands Region experienced rapid growth from 1997 to 2001, changes to RAMP were made annually. These changes not only affected RAMP's objectives, and organizational structure, but the study area and study design as well. Potential sampling methods, sentinel species and reference lakes and streams were also evaluated during this period. Some methods were adopted and then abandoned during the program.

This is a review primarily of the biostatistical analysis conducted as part of this first five years of the program.

The entire Five Year Report was reviewed to examine if the analyses conducted in the report are suitable, if the conclusions can be supported by the analyses chosen, and to make recommendations for changes to future years of RAMP. A less detailed review of the interim reports was also conducted (Appendix IV).

2. GENERAL COMMENTS

2.1 Replication and pseudo-replication.

A major concern in Environmental Impact studies is proper replication and the avoidance of pseudo-replication (Hurlburt, 1984). Replication provides information about the variability of the collected data under identical treatment conditions so that differences among treatments can be compared to variation within treatments. This is the fundamental principle of ANOVA.

For example, consider a survey to investigate sediment quality at various locations on a river. A simple design may take a single sample at each of 4 locations:

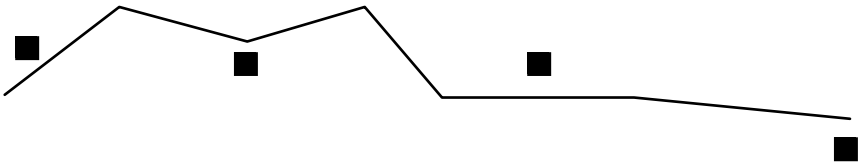


Figure 2.1(a) A simple survey that provides little useful information.

These four values are insufficient for any comparison of the variable across the four locations because the natural variation present in readings at a particular location is not known.

In many ecological field studies, the concepts of experimental units and randomization of treatments to experimental units are not directly applicable

making “replication” somewhat problematic. Replication is consequently defined as the taking of multiple INDEPENDENT samples from a particular location. The replicated samples should be located sufficiently far from the first location so that local influences that are site specific do not operate in common on the two samples. The exact distance between samples depends upon the biological process. For example if the locations are tens of kilometers apart, then spacing the samples hundreds of meters apart will likely do for most situations. This gives rise to the following design:

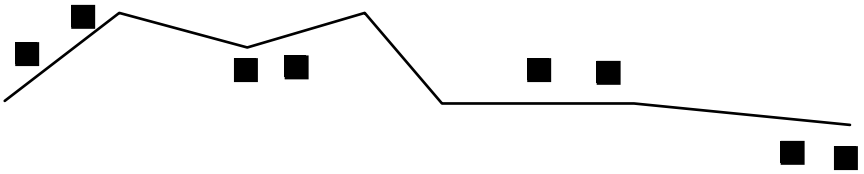


Figure 2.1(b). A replicated survey (if the points are independent within a pair).

Now a statistical comparison can be performed to investigate if the mean response is equal at four locations. This particular design would give rise to the following statistical model

Y = location sample(location)-R

where *location* represents the effect of different locations, and *sample(location)-R* represents the random, independent replicates at each location. The ANOVA table would construct a test for location effects using the F-ratio of

$$F = \frac{ms(location)}{ms(sample(location))}$$

with the idea that variation in means among locations would be compared to variation in readings within a location.

The key point is that the samples should be independent but still representative of that particular location. Hence, taking two samples from the exact same location, or splitting the sample in two and doing two analyses on the split sample will not provide true replication. These would be pseudo-replicates. Hurlburt (1984) defines pseudo-replication as

> “Pseudo-replication is defined as the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent.”

Consequently, a design where duplicate samples or split-samples are taken from the exact same location (Figure 2.1(c)) would be an example of pseudo-replication.

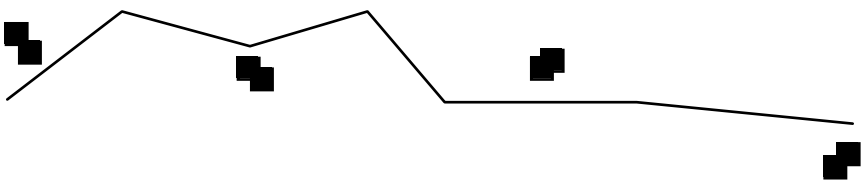


Figure 2.1(c). A pseudo-replicated survey.

Note that the data from Figure 2.1(b) and Figure 2.1(c) looks “identical”, i.e. pairs of “replicated” observations from four locations. Consequently, it would be very tempting to analyze both experiments using exactly the same statistical model and ANOVA table. However, there is a major difference in interpretation of the results.

The design in Figure 2.1(b) with real replicates enables statements to be made about differences in the mean response among those four general locations. However, the design in Figure 2.1(c) with pseudo-replication only allows statements to be made about differences among those four particular sampling sites which may not truly reflect differences among the broader locations.

Obviously the line between real and pseudo-replication is somewhat ill-defined. Exactly how far apart do sampling sites have to be before they can be considered to be independent. There is no hard and fast rule and biological consideration and knowledge of the processes involved in the environmental impact must be used to make a judgment call.

The same considerations apply when sampling across time. Samples need to be taken far enough apart in time so that they are independent. For example, if data from continuous logging is used (say every minute over a year), then it would be unfair to treat all 500,000+ observations are being independent when a regression line is fit.

What is the relevance to the RAMP report? In some part of the report, this has been recognized. For example, Section 6.1.1 (page 6-25) states:

> “Individual samples collected from the same site do not represent replicates in the statistical sense because they are not independent. Widely-spaced samples from a reach (each sample representing a site) were used as replicates to compare reaches.”

But, consider Section 4 of the report and Figures 4.12 and 4.13. The authors again recognize some obvious pseudo-replication (e.g. only one measurement is selected from multiple measurements in a location in a day), but Figures 4.12 and 4.13 show clustering of data points at a larger time scale. Hence treating all the points in these figures as independent likely overstates the observed relationship, i.e. the reported p-value is too small. Other cases of potential pseudo-replication are cited in the Technical Comments below.

Another consequence of pseudo-replication is that estimates of variation used in power analyses are too small which lead to underestimates of the required sample size to detect a specified difference.

All of the analyses in the report should be reviewed with the dangers in pseudo-replication in mind. The report should also provide a clearer description of the sampling

design – the text at the bottom of page 6-5 could serve as a prototype for similar statements in the other chapters of the report.

2.2 Matching Analysis with Design

Another common concern with environmental field studies is ensuring that the analysis matches the design by which the data were collected. All two-factor designs are not analyzed in the same way!

For example, consider (as in Chapter 5) a study to compare a variable in sediments among four locations and two sides of the river[5].

Two possible design are shown in Figures 2.2(a) and 2.2(b)

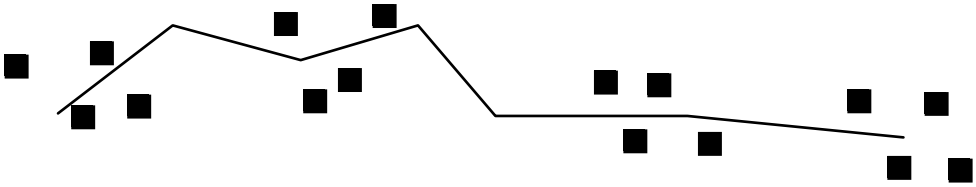


Figure 2.2(a) A design to study the effects of side and location with independent replicates at each location/side combination.

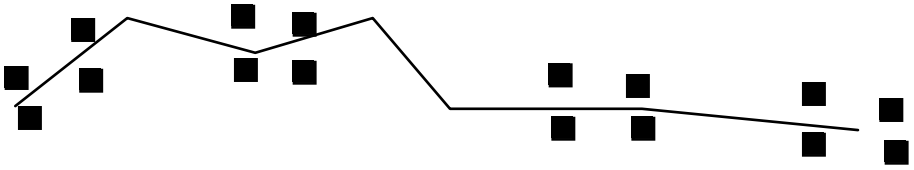


Figure 2.2(b) A design to study the effects of side and location with paired replicated at each location/size combination.

Both designs have exactly the same number of data points and without looking carefully at how the data were collected, the raw data does not provide information about the actual design. However, the analysis of these designs is quite different.

The analysis of the design in Figure 2.2(a) would use the statistical model

Y = location side location*side sample(location*side)-R

where *location*, *side*, and *location*side* represent the effects of location, side, and their potential interaction (i.e. is the effect of side consistent among all locations?). The term *sample(location*side)-R* represents the (random) variation of the response variable among replicate samples at the same location/side. Because there are independent replicates at each side/location, an model-independent estimate of this variation is available directly from the data. Note that this model makes an implicit assumption is that

[5] It is not necessary to take replicate samples at ALL side-location combinations, nor is it necessary to have equal number of replication samples at ALL side-location combinations. However, balanced designs (with equal number of replicates) have the advantage that tests for each effect are now orthogonal to each other and that simple software can be used to analyze these designs.

the replicate samples on each side of the river are independent among themselves on the side and among the two sides of the river. Hence the sample points are take NOT directly across from each other on both sides of the river. The statistical comparisons would be computed as:

$$F_{location} = \frac{ms(location)}{ms(sample(l*s))}, F_{side} = \frac{ms(side)}{ms(sample(l*s))}, F_{\text{interaction}} = \frac{ms(\text{interaction})}{ms(sample(l*s))}$$

However, the analysis of the design in Figure 2.2(b) must now allow for the existence of potential small scale effects within each location that affect both sides of the river simultaneously? This would render the two replicate samples on each side no-longer independent. There are two "sizes" of effects. First location effects operate on a large scale (on sets of 4 samples) while micro-location effects operate on paired points on each side of the stream. The statistical model is now:

Y = location site(location)-R side location*side residual-R

where *location*, *side*, and *location*side* terms again represent the effects of location, side, and their potential interaction. The *site(location)-R* term represent the micro-location effects that affects both sides of the river simultaneously. Because there are replicate pairs of points at each location, the within location variation can be computed independently of the model. The *residual*-R term represents the variation among individual sample points and is found by subtraction.[6] This model is a variant of a split-plot design with locations being main plots, and the sides of the river within each pair at each location being the subplots. The

$$F_{location} = \frac{ms(location)}{ms(site(location))}, F_{side} = \frac{ms(side)}{ms(residual)}, F_{\text{interaction}} = \frac{ms(\text{interaction})}{ms(residual)}$$

Notice that the test for location is NO LONGER computed using the residual variation – it must be constructed using the site-to-site variation within each location. The reason for this is that there are now two scales of effects – location effects affect groups of 4 points, while the site effects within location affect a pair of points (both sides of the river).

The situation becomes more complex once sampling is replicated across years. Again, consider the first design where the sampling is repeated in two years:

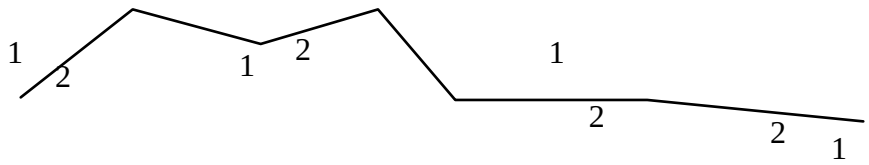


Figure 2.2(c) An (inadequate) sampling design with independent measurements across time. The values 1/2 represent years 1/2 measurements

Here, the sample points in year 2 are situated at random within each location ignoring bank effects but far enough from the original sample location to be independent of micro-location effects. [This design suffers from the same defect as outlined earlier,

[6] This design could have replicate points at each side within each pair at each location which would then allow a model-independent estimate of this variation to also be computed.

i.e. no real replication, but is used only to illustrate a comparison with a paired design below.] Contrast this to the design in Figure 2.2(d) where sampling is deliberately located at the SAME sampling sites in both years:

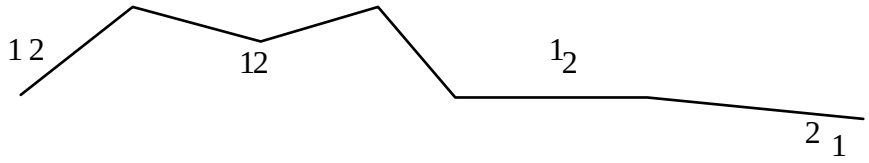


Figure 2.2(d) An (inadequate) sampling design with paired measurements across time. The values 1/2 represent years 1/2 measurements

Now a proper model that incorporates time effect must account for both the large scale location effects but also potential micro-location effects. [Again, there is no replication at any of the location –year points and so is a poor design.]

This is relevant to the RAMP report because many of the chapters involve two (or more) factor models but the reports always treat the data as if it came from completely randomized designs (as in Figure 2.2(a)) rather than looking closely at how the data were collected. In many cases, time is a factor, and it is not clear if sample points are paired across time or are independent across time. The analysis is different in these two cases. The report should pay more attention to how the data were collected

Additional examples are provided in Morrison et al (2001) on the need for proper matching of design and analysis.

2.3 Lack of suitable replication - consequences

In many cases, it appears that no suitable replication was collected during the sampling design. Rather than simply throwing up ones hands and abandoning the analysis, what are the consequences of no real replicates?

Consider again (as in Chapter 5) a two factor design to investigate the effects of location and bank upon sediment quality. A simple design might take samples from each side of the bank at each of the 4 locations:

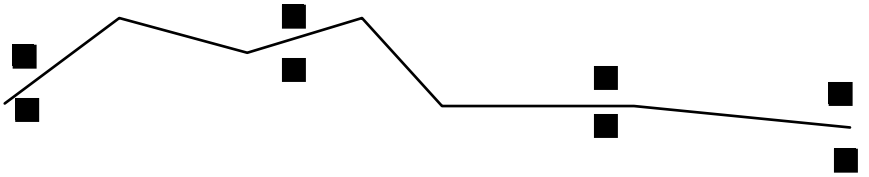


Figure 2.3(a) A design to compare effects of bank and location without replicates.

At first glance, this appears to be similar to the previous designs (Figure 2.1(b) or Figure 2.1(c)) with the same number of total replicates except they are now taken on both sides of the river. However, unless a very strong, untestable assumption is made, no valid statistical test can be made for the effect of side or the effect of location! This design has NO real replicates.

The assumption that must be made is that the effects of side is EQUAL at all locations and that the effect of location is EQUAL on both sides, i.e. that there is NO interaction between the effects of location or site. The is insufficient information in the experiment to test this assumption.[7] The statistical model for this experiment under this very strong assumption would be

Y = location side residual-R

where *location* and *side* represent the effects of side and location respectively. The *residual-R* term represents the residual, random variation, after adjusting for location and side. Note that unlike the previous model, the *residual-R* term can only be computed after adjusting for side and location – there is no data-driven estimate of this variation.

This model appears to be the same as the model as for a randomized block design. [A key assumption of a randomized block design is that blocks and factors also do not interact]. However, there is subtle difference between factors and blocks that will not be discussed in this report that implies that they are not identical. The F-statistics for testing effects of location or side would be computed as:

$$F_{location} = \frac{ms(location)}{ms(residual)}, F_{side} = \frac{ms(side)}{ms(residual)}$$

where *ms(residual)* represents the remainder after adjusting for the effects of *side* and *location.*

So on first glance, it does appear that a valid statistical test has been performed – but it will only be valid if the assumption of no interaction is true.

The situation becomes more complex once sampling is replicated across years! Again, consider the first design where the sampling is repeated in two years:

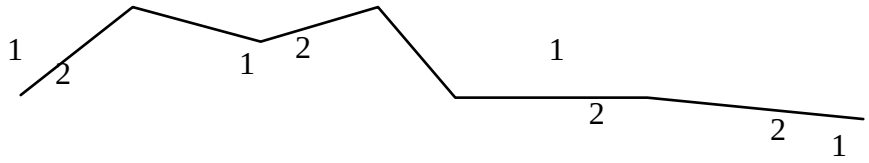


Figure 2.3(b). A design to compare the effects of location and time without replication and independent randomizations in each year.

Here, the sample points in year 2 are situated at random within each location ignoring bank effects but far enough from the original sample location to be independent of micro-location effects. This design suffers from the same defect as outlined earlier, i.e. no real replication and so analyses are only possible if a very strong, untestable assumption is made – namely, no year*location interaction, i.e. the year effects are equal for all locations, and the location effects are equal for all years. The model that must be fit is

[7] A crude profile plot of the value of the response variable at each location for each side could be used to informally check if the profiles are parallel which would indicate that no apparent interaction exists, but this is only an informal check.

Y = location year residual-R

where *location* and *year* represent the location and year effects and *residual-R* represents the random variation that must be found after fitting the model. The design can again be improved by replicating the measurements at location-year combinations.

Contrast this to the design where sampling is deliberately located at the SAME sampling sites in both years:

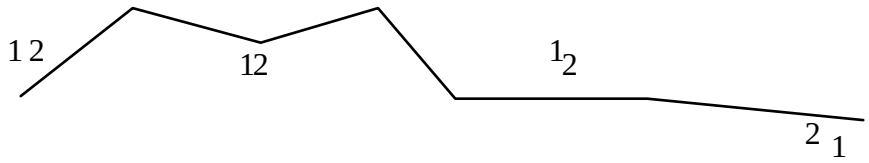


Figure 2.3(c) A design to compare the effects of location and time with paired observations across years.

Now a proper model must account for both the large scale location effects but also potential micro-location effects. However, because there is no replication at any of the location –year points, a model can only be fit if strong untestable assumptions are made – in particular that there is no year-location interaction and that there is NO micro-location effect. The model for this design is:

Y = location year residual-R

This is the same model as for the previous case, but this is an artifact of the poor experimental design chosen – without proper replication, only very simple models that make strong assumptions can be fit. There is a fundamental difference between these two designs – the former is akin to a completely randomized design while the latter is a variant of a split-plot design. With proper replication the model would look quite different.

The RAMP report has many comparisons of the above type. In general, these comparisons may be misleading because of the lack of proper replication. At the very least, these implicit assumptions should be stated directly in the report.

2.4 Reporting results; p-values and power analyses

The report has numerous tables reporting the results of testing for the effects of various factors. In many cases, p-values are the statistic of choice and in some cases, only an indication of statistical significance (i.e. $p<.05$) are provided.

Many authors have reviewed the problems with p-values (e.g. Steidl et al 1997; Cherry, 1998; Johnson, 1999). Basically, the p-value does not provide sufficient information to assess the magnitude of the difference detected and can be misleading to readers. Other problems include:

- The choice of null hypothesis is often arbitrary.
- Conclusions are stated a rejecting or not-rejecting the hypothesis when in fact the data may not be that clear cut.
- The choice of α-level (i.e. 0.05 significance level) is arbitrary. Should

different decisions be made if the p-value is 0.0499 or 0.0501? The value of α used in a study should reflect the costs of Type I errors, i.e. the costs of false positive results and the costs of Type II errors, i.e. the costs of false negative results.

- Users of statistics have often emphasized certain standard levels of significance such as 10%, 5%, or 1% indicated (typically) by asterisks. These reflect a time when it was quite impossible to compute the exact p-values, and only tables were available. In this modern era, there is no excuse for failing to report the exact p-value.
- In many cases, hypothesis testing is used when the evidence is obvious. This leads to statements similar to "p<.00001".
- P-values are prone to mis-interpretation as they measure the plausibility of the data assuming the null hypothsis is true, rather than measuring the "truthfulness" of the hypothesis.
- P-values are highly affected by sample size. With sufficiently large sample sizes every effect is statistically significant but may be of no biological interest.
- The tradeoffs between Type I and II errors, power, and sample size are rarely discussed in this context.
- Just because the null hypothesis is rejected does not imply that the effect is very large. For example, if you were to test if a coin were fair and were able to toss it 1,000,000 times, you would reject the null hypothesis of fairness if the observed proportion of heads was 50.001%. But for all intents and purposes, the coin is fair enough for real use. Statistical significance is not the same as practical significance. Other examples of this trap, are the numerous studies that show cancerous effects of certain foods. Unfortunately, the estimated increase in risk from these studies is often less than 1/100 of 1%!
- Just because an experiment fails to reject the null hypothesis, does not mean that there is no effect! A Type II error - a false negative error - may have been committed. These usually occur when experiments are too small (i.e. inadequate sample size) to detect effects of interest.
- In some experiments, hundreds of statistical tests are performed. However, remember that the p-value represents the chance that this data could have occurred given that the hypothesis is true. So a p-value of 0.01 implies, that this event could have occurred in about 1% of cases EVEN IF THE NULL IS TRUE. So finding one or two significant results out of hundreds of tests is not surprising!

Some of the problems with p-values were recognized in the report. For example, page 7-95 states:

> "In many studies, a statistically significant difference in biological measures is used as evidence that a change has occurred. Indeed, several industry-wide monitoring programs have adopted this approach (Environment Canada 1998, 2002). Unfortunately, extrapolation from statistical significance to ecological

significance is difficult because statistical significance depends upon sample size, and may not relate to the size of the impact."

The report recommends that ecological significance be stated in terms of the variability of the natural populations:

> "The approach proposed by Kilgour et al. (1998) was used to determine the ecological significance of the observed differences. They define ecologicallyrelevant differences as observations from impact locations that fall outside the normal range of variation based on reference-location data. They also define the normal range as the region enclosing 95% of reference-location observations. The 95% region can then be expressed generically as standard deviations in univariate responses. For example, in single responses that are normally distributed, the region defined by $\mu \pm 1\ \sigma$ incorporates about 67% of the population, and $\mu \pm 1.96\ \sigma$ incorporates about 95% of the population. These calculations were performed with the RAMP data, and all mean values of exposure population parameters fell within the normal range based on the three reference populations; …"

While this an improvement over the lack of determination of an ecologically significant result, the report should review these proposed ecologically significant effects carefully because changes in the mean that are much smaller than a standard deviation of individual observations can have large ecological impacts.

Rather than relying upon p-values a summary measure, the earlier cited papers have suggested that more emphasis be place on confidence intervals for effect sizes. For example, the report contains many tables such as Table 4.17, where a comparison between two levels of a factor is shown. The mean values for each level are shown, and the F-statistic is shown with asterisks (* or ** or ***) representing if the effect was "significant". These types of tables could be greatly improved if the estimated difference was shown along with the estimated confidence interval for the difference. In this way, the reader can assess the magnitude of the differences and if these are biologically important.

Along with reporting confidence intervals for effect sizes, power analyses provide information on the likelihood of success in detecting real changes. The report has numerous power computations, but these could be improved/corrected in the following ways:

- better terminology, e.g. "effect size" should read "minimum detectable difference"
- using the proper estimate of variation. As noted below, pseudo-replication will lead to estimates of variation that are too small and estimates of minimum detectable differences that are too small, i.e. the actual power is much less than "advertised".
- discussion of power of trend tests confuse the minimum sample size that is technically needed to compute a statistic with the sample size needed to detect a

specified decline. For example, the Mann-Kendall test looks for monotonicity of the trend and this is dependent upon the actual slope and the natural variation in addition to the sample size. The first two components were not discussed at all in the report.

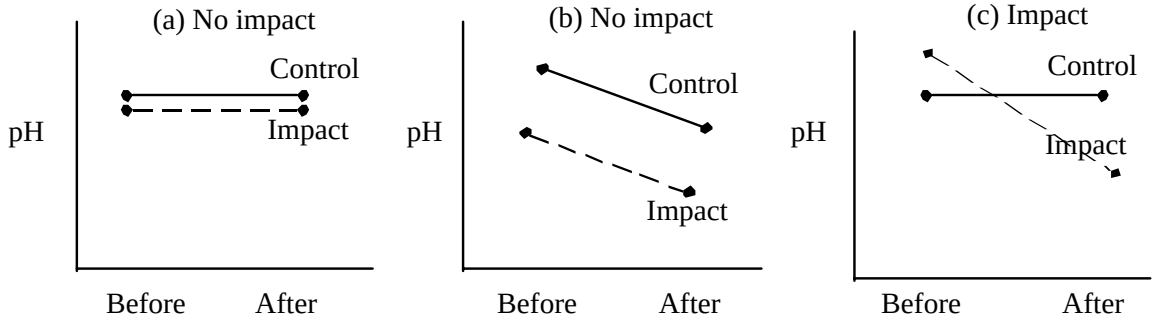
2.5 Types of Monitoring Designs

The simplest monitoring design is a before/after measurement at a single site. For example, soil pH is measured before and after emissions begin. This design is widely used in response to obvious accidental incidences of potential impact (e.g. oil spills, forest fires), where, fortuitously, some prior information is available. In these types of studies, the manager obtains a single measurement of pH before and after the event. If the second survey reveals a change, this is attributed to the event.

Unfortunately, there may be no relationship between the observed event and the changes in the pH - the change may be entirely coincidental. Even worse, there is no information collected on the natural variability of the pH over time and the observed changes may simply be due to natural fluctuations over time. Decisions based on this design are extremely hard to justify.

The most basic monitoring design that can distinguish natural changes from changes that follow an impact are the Before/After/Control/Impact (BACI) design where pH is measured at the site before and after the impact, and at a control site (not affected by the impact) before and after the impact. Figure 2.5.1 illustrates three possible outcomes.

Figure 2.5.1.Examples of outcomes from a BACI Impact Study



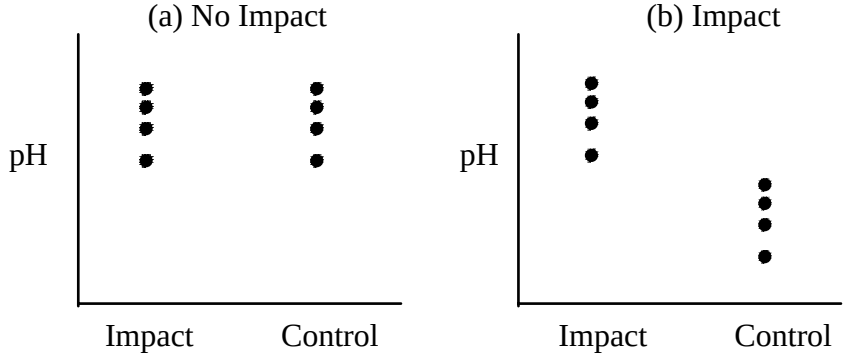
In Figure 2.5.1(a), the pH measurements did not change from before to after the impact at either the control or impacted site and there is no evidence of an impact. In Figure 2.5.1(b), both sites have changes in pH over time, but the change is equal for both sites. Because both sites changed in a parallel fashion, there is no evidence of a differential effect of the impact. In Figure 2.5.1(c), the change is no longer parallel between both sites, and there is evidence of an impact.

But what can be done if baseline (before impact) measurements are available.

Weins and Parker (1995) considered the problem of assessing environmental impacts when before measurements are not possible (e.g. in the case of accidental impacts such as oil spills). They divide potential designs into various classes – the two most relevant to this study are the single-time designs (the current document) and the multiple-time designs (for the future).

In the absence of before measurements, the single-time designs take samples from several sites within the impacted area and from several sites outside the impacted areas. For example, Figure 2.5.2 illustrates this design with four sites chosen from the impacted area and four sites chosen from the control area.

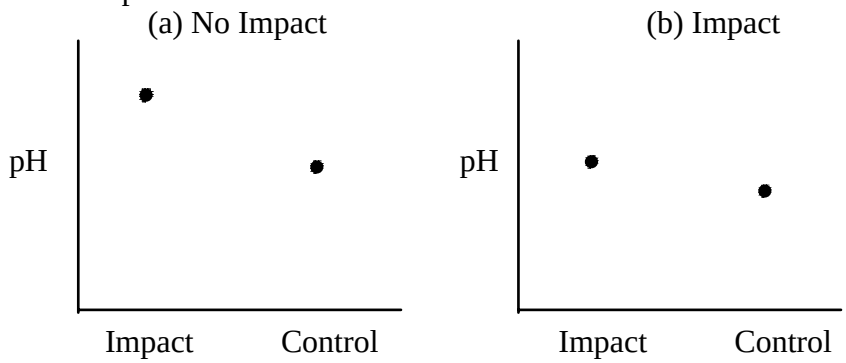
Figure 2.5.2: Example of a single-time design



In Figure 2.5.2(a), the average pH level is about the same in both areas while in Figure 2.5.2(b), there is clear evidence that the mean has declined. The greatest danger with this design is that the observed difference between impacted and control sites may just be due to random variation and not related to the impact but carefully choosing control areas to be a similar as possible to the impacted areas should reduce this possibility.

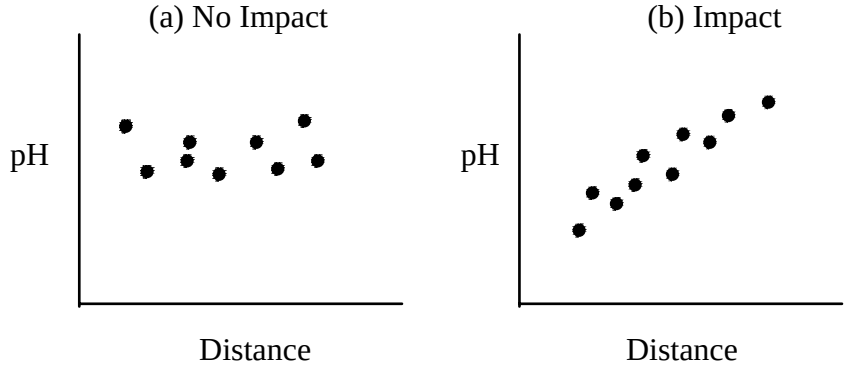
Note that replication within the impact and control areas is also vital as mentioned earlier. As an illustration of the danger that no replication poses, consider Figure 2.5.3 - the same values are used as in Figure 2.5.2, except that only one site was measured from each area. Just by chance, these happened to correspond to the highest and lowest pH in each group.

Figure 2.5.3: Example of misleading results that could be obtained without replication.



In some cases, such in this study, the impact can be quantified and a gradient of exposure can be established. For example, distance from the source of emissions may be used as a surrogate for exposure. Figure 2.5.4 illustrates two potential relationships:

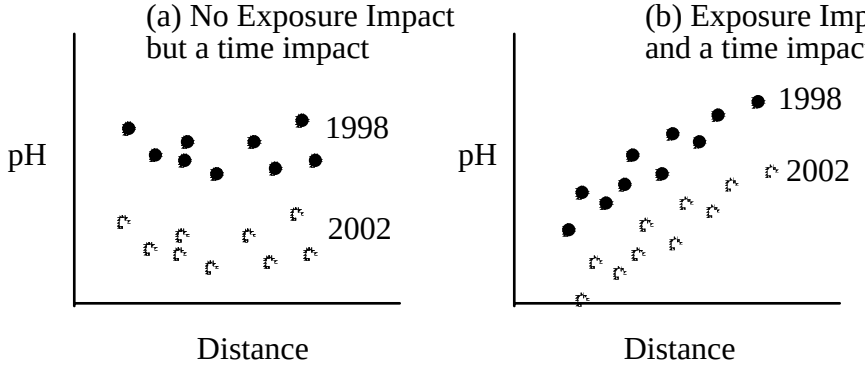
Figure 2.5.4: Example single-time design with a gradient of exposure



Note that a wide range of exposures needs to be monitored and that the design assumes that all other factors that might affect the response are equal except for exposure. For example, it may turn out that all sites are located in a northerly direction from the emission source and latitude effects are what causing the response variable to change.

This study is intended to continue over time, and so both temporal and exposure effects can be examined as outlined in Weins and Parker (1995). Figure 2.5.5 illustrates two responses that could occur (others are possible):

Figure 2.5.5: Example of a multiple-time design study with gradient of exposure



In Figure 2.5.5(a), there does not appear to be any relationship of the response to exposure but something appears to be happening over time. In Figure 2.5.5(b), there appears to be a relationship with exposure and again some effects of time.

The lack of baseline information for some aspects of the study have been recognized by the authors. For example in section 4.3.3, the report states:

> "In response to the question "Is RAMP collecting or otherwise obtaining the type of information required to differentiate natural variability from changes associated with human activity?", the answer is mixed. In the case of sulphate levels in the Muskeg River, adequate baseline data had been collected before and after the initiation of development at both upstream and downstream locations to clearly identify a significant change attributable to human activity in the basin. However, as discussed in Section 4.3.3, sufficient baseline information may not be available in less well-studied systems to determine if, for example, significant temporal variations can be detected prior to development.

The Weins and Parker (1995) paper should be reviewed to see if their suggested designs may provide further monitoring options for this study. This is alluded to in the report:

> "Since there will be a potential for the appearance of long-term trends unrelated to oil sands developments (e.g., due to climate change or long-term hydrological cycles), monitoring to detect long-term trends should incorporate at least one reference river. Although the analysis described in Section 6.2.1.2 suggests that each river is unique in terms of its benthic community, it is possible that long term trends unrelated to development would be similar in all regional rivers. This would allow the consideration of time-trends observed in reference rivers in the interpretation of data from potentially affected rivers. Based on the extent of planned oil sands development in the region and its hydrological features, finding reference rivers is problematic. Therefore, if significant long-term trends are found by future assessments without corresponding reference river data, the possibility of factors other than oil sands developments causing the observed trends will need to be considered, possibly by evaluating the consistency of trends among rivers monitored throughout the region."

A more systematic exploration of potential monitoring designs for this case should be included in the report.

Finally, the report comments many times that the same monitoring station was not measured over time or that stations are added or dropped over time. There are obvious tradeoffs between fixed monitoring stations and random monitoring stations which are discussed in many books. However, the RAMP steering committee should consider using **panel-designs** which are a combination of fixed and random monitoring stations. A classical panel design would, for example, start with 12 monitoring stations, and allow up to 1/3 of the stations to rotate each year. Some stations, if feasible, could not be rotated. A simple example is shown below:

	Stations monitored												
yr	1	2	3	4	5	6	7	8	9	10	11	12	13
1	x	x	x	x	x	x	x						
2	x	x	x	x	x	x	x	x	x	x			
3	x	x	x	x	x	x	x	x	x	x	x	x	x
4	x	x	x	x	x	x	x	x	x	x	x	x	x

5	x	x	x	x			x	x	x	x	x	x
6	x	x	x	x						x	x	x

These designs combined the best features of fixed and random monitoring. A few stations have long term measurements, while the rotated stations allow for sample refreshment (because, for example, of natural disaster at a station or change in conditions at the station).

2.6 Data storage, meta data

This topic is missing from the report, but the RAMP review should look at how the data for this study is being stored. In particular, the use of simple Excel-type spreadsheets may be inadequate as linking between sheets of different information from the same location may be missing.

An important component of data storage is consideration of maintenance of meta-data, i.e. information about the actual data such a location, sampling method, who collected, who analyzed the data etc. How is this information being stored?

2.7 Choice of analysis methods.

The report uses three basic methods.

2.7.1 Estimation of extreme values (precipitation, stream flow, and temperature)
The report uses two programs – the Consolidated Frequency Analysis (CFA) from Environment Canada and the FRQ from Kite to estimate return events. Based upon a reading of the report, these appear to be appropriate methodologies.

However, the described methods of analysis in Section 3.2.1.1 (Precipitation events) is rather unclear. Unlike annual min/maximum records, there is only value per year for precipitation. Consequently, how is the data separated into wet and dry years prior to fitting the appropriate extreme value distribution? Different separations would lead to different estimates of wet/dry return periods.

Note that the precision of the estimated events is lively very poor. Chow (1977) wrote that in order to accurately predict a 10-year recurrence frequency event, 100 years of records are needed, but, in order to accurately predict a 100-year recurrence event, about1,000 years of records will be needed.

The report did not do a power analysis to examine the size of changes that can be detected given the available data. I suspect that the power is very poor given the extreme variation in the data, so that it may not be cost effective to even monitor these variables.

Consequently, I would suggest that (a) the estimated parameters of the fitted distribution should be displayed in tables such as Table 3.10, so that future users do not need to refit the data and (b) a proper power analysis be done.

2.7.2 ANOVA

The report uses ANOVA extensively to investigate if changes in the mean response have occurred among locations, year, bank, upstream/downstream etc.

ANOVA is a very general methodology but it is extremely important that designs have proper replication (see earlier comments) and that the analysis matches the experimental design. As seen in my comments below, there are several instances where pseudo-replication is apparently taking place, where no real replication occurs, or where the wrong model has been fit.

Some of the analyses need to be redone using the appropriate replicates and/or models.

The report commonly reports p-values but does not report estimates of effect sizes. As noted earlier, it is better practice to report effect sizes rather than simple p-values which have a number of "defects".

Power analyses may need to be redone to incorporate the proper estimate of variation, i.e. pseudo-replication typically leads to estimates of variation that are too small and power estimates that are too large; in split-plot designs the different error terms are used for power analyses of the different factors.

2.7.3 Regression

The report uses regression analysis to check for temporal trends. Often a non-parametric regression method (Mann-Kendall method) is used.

The primary concern that I have with the regression analyses have to do with the failure of the observations to be independent, e.g. pseudo-replicates are used as real-replicates.

The report attempts to do a power analysis for the non-parametric testing method but confuses the technical minimum sample sizes with a the real power to detect a specified trend. These should be redone to properly report the power to detect, e.g. a 10% decline over 10 years.

2.7.4 Principal Component Analysis

Principal Component Analysis (PCA) is commonly used to reduce a large set of inter-correlated variables to a smaller set of underlying "variables". For example, in Chapter 4, PCA was used to reduce a large number of water chemistry variables to a smaller set – for example, many metals seem to vary together among the samples.

The idea of PCA is to extract a component that has the highest possible variance; then extract a second component that has the next highest variance but is orthogonal to the first etc. PCA is common done on the correlation matrix of the observations as the correlation matrix does not change if the measuring scale of individual variables changes.

For example, changing the measuring units from cm to m would reduce the variance of a variable by a factor of 100, but has not effect on its correlation with other variables. I was somewhat puzzled then by the analysis in Chapter 6 where the PCA was done on the covariance matrix which is not measuring-scale independent.

The interpretation of a principal component (PC) is obtained by examining the correlation of each individual variable with the extracted PC. Sets of variables that are highly correlated provide an interpretation of the component. For example in Table 4.9, a reasonable interpretation of the first component is "concentration of heavy metals", while that for the second components appears to be related to "salts". The report appears to have interpreted the extracted PC correctly.

The extracted PC are often then regressed against other variables, e.g. stream flow. Again, the report appears to have done these appropriately.

The usefulness of PCA for environmental impact studies is mixed. A PCA could be used to identify a common factor that may be easier to measure than a set of disparate variables. However, in some cases, there is little to be gained – for example, chemical analyses of water use methods that produce the individual constituent components a very little marginal cost. When additional data are collected and included in a PCA analysis, the new data can change the computation of the PC slightly so results cannot be directly compared across years. It would be advantageous to use the PCA results to define a new variable (e.g. sum of total metal concentrations) whose definition does not change from year to year.

3 DETAILED TECHNICAL COMMENTS

3.1 Chapter 1 comments

Section 1.7.4 Changes in monitor plans over time – consider a panel design with sites being rotated in and out a suitable design?

3.2 Chapter 3 comments

a) Section 3.2.1.1 - Precipitation methods

Need to carefully define the calendar year. For example, snowfall is recorded from October to May which crosses a year boundary. To which year is this assigned? For example in Table 3.6, calendar year 1945 has both snowfall and rainfall records? The report needs an exact definition of the recording year., e.g. 1945 year corresponds to Sept 1 1944-31 August 1945.

How can two different distributions be fit for dry and wet years to determine return periods for 10/100 year events? It seems to me that a single curve needs to be fit to the entire data and the appropriate percentiles determined (e.g. the .01, .1 or .9 and .99 points.

b) Section 3.2.2.1 - Temperature methods

Same comments as above. Dec-Feb data belong to which year? For example in Table 3.8, is Dec-Feb for 1945 associated with Dec 1944-Feb 1945 or Dec 1945-Feb 1946.

Second last paragraph with "correlation between two data sets" and "… cold winter unlikely to CAUSE a cold summer…". Correlation does not imply causation.

c) Section 3.2.3.1 - Runoff depth analysis

The consolidated frequency analysis program of Environment Canada and the FRQ program by Kite were used to runoff depth – were these used for previous two sections on Precipitation and Temperature?

Figure 3.12 legend differs from rest of graphs in series. Try and make all legends/axes/lines consistent. Try and use consistent colors through out the document, e.g. for the 100/10 return periods values.

Why is goodness of fit used to select appropriate distribution? More modern theory would use AIC for model selection and model averaging (Burnham and Anderson, 2002). Using the distribution that fits best, may lead to a better fit than can be justified..

Report the fitted parameters of distributions so others can use these to estimate other return periods etc without having to reaccess the raw data.

d) Section 3.3 Testing for temporal trends.

Estimate the actual trend line and report a 95% confidence interval – this can be done even with non-parametric methods as done in later chapters of the report. Absence of a detection of a trend does not imply that there is no trend – rather that it may be small relative to the effect size. Show a plot with the fitted trend curve.

If you find serial dependence in the data set, does it make sense to do Spearman test for trend which assumes independent data points? I suspect that this non-independence makes the Spearman test incorrect – an example of pseudo-replication.

Report the actual p-value of the test statistics rather than simple if significant at the 1% or 5% levels. Report confidence intervals when ever possible. See the papers cited in the introduction on problems with the way the results are presented.

e) Section 3.3.1.1 – Precipitation – Results and Discussion

"Difference at the 95% confidence interval" makes no sense. Reword here and elsewhere in document.

f) Section 3.3.1.2 - Testing for trend in temperature

The split-sample test – was the year 1971 specified a-priori or was data snooping used?

g) Section 3.4 Monitoring To Verify EIA Predictions

What is missing is a estimate of the size of change than can be detected given the monitoring data, i.e. a power analysis of the recorded data. I suspect that this is extremely low given the high variability in the data.

Consequently, it may not be sensible to collect this data if it has essentially no chance of detecting any reasonable size of impact!

Figure 3.50 looks strange as lines are not parallel to X-axis.

3.3 Chapter 4 comments

h) Section 4.1.1 – Program Overview

Because not all sites sampled in all years, think of a panel design.

There is a big discrepancy in sample sizes among tables. Did Athasabasca really have 300-800 samples or do some of these include split/duplicate samples? Look like the potential for lots of pseudo-replication.

i) 4.2.1.1 - Methods

"Split and duplicate samples were reduced to single samples to guarantee data independence. This process was completed through either random selection or, in cases of unequal analysis, by choosing the sample that had been submitted for the more complete analysis."

While the goal of achieving independence among the samples is laudable, the approach is crude and may "waste" information. Duplicate/split samples are easily handled in modern statistical software through nesting terms. At the very least, the average of the split/duplicate samples should have been used rather than using a single random selection.

"… values recorded as zero were eliminated". Is this really true? A zero value is NOT the same a not recorded and contains valuable information. I suspect this was to avoid problems with log(0) in the analyses, but why is real data eliminated? It is not clear in the remainder of this section if 0 values were excluded for all analyses.

I can see eliminating entire class of variables if the majority of readings are non-detectable, but this also has dangers. For example, suppose that upstream of a oilsand project a certain component is non-detectible, but downstream from an oilsand project, most a non-detectable but, around 20% show extreme levels of a chemical?

The method of dealing with non-detectible (assigning 1/2 of the nd limit) is crude, but will work reasonable well as long as the number of nd is relative small (say 20% of the dataset or less).

j) Section 4.2.1.1.- Explicit TSS relationship

Do the plots first to see any outliers or weird points that may reduce the sample correlation coefficient to zero regardless if a linear relationship exists or not.

Note the problem with p-values. A correlation in Table 410 of .33 was significant for the Athabasca River but not for the Wetlands solely because of sample sizes of around 300 and 30 in the two locations.

k) Section 4.2.2.1 - Methods

Use ANCOVA to see if relationship is the same between the different sources. Is this possible as the flow variable is different in different streams?

l) Section 4.2.2.2 – Results and Discussion

Analysis pools over all years/seasons. A more complex model should be used to account for year/season effects.

m) Section 4.2.3.2 – Results and Discussion.

Analysis is incorrect. Replicate measurements within a season are pseudo-replicates (Hurlbert, 1984) and cannot be treated as independent sample points. A model such as

Y = year season year*season sample(year*season)

should be fit so that the test for season is against the year*season interaction, or an "average" must be computed for each season to give ONE measurement per year/season combination. The consequences of the incorrect analysis in the report is typically too many significant results.

As pointed out, many of the variables of interest are highly related to stream flow which is also related to season. Hence, the test of season is essentially a test of stream-flow.

Earlier, analyses were done on log(concentrations), but this section's writing makes it sound like the analyses were done on the raw concentrations. For example in Table 4.15 – Table 4.16 report what appear to be simple MEANS rather than geometric means if the analysis is done on the log-scale?

n) Section 4.3.1.2 – Analysis of Temporal trends in water quality

It appears that analysis is incorrect. There are multiple measurements taken in any particular year that are likely highly correlated, but these are treated as independent observations. For example, there are only about 25 years of data in the long-term study, but over 150 data points are presented. This can be seen in Figures 4.12 and 4.13 where there is evident clustering of points within years. Again, the likely effect is too many significant results.

Similar comments about the analysis in Tables 4.20 and 4.21 – the sample sizes are not real, independent measurements but are pseudo-replicates. In the Section on the Muskeg River, the report suggests the continuous measurements increase the sample sizes. Again this is pseudo-replication – two measurements taken very close together in time and space are not the same as two independent measurements.

4.3.1.3 – Conclusions and Recommendation.
Absence of a statistically significant results does not imply the non-existence of an effect. Power and sample size may have been inadequate to detect a change of biological importance.

4.3.2.1. –Trends in Athabasca River
See earlier comments in Section 4.3.1.2 – I suspect the analyses are incorrect because of non-independence of the data. Indeed Table 4.22 shows sample sizes that are too large for the model fit – i.e. include pseudo-replication within the year/season/location terms of the model. If year and season are blocking variables, then the year*season interaction term should be included to make this a paired design for testing location. Not a BACI design, so even if differences are detected, these cannot be attributed to oil sand development, but rather may have always existed.

4.3.2.1 – Trends in Muskeg River
This is a BACI design as pre-development data is available. Same problems as before with pseudo-replication within each year/location/season combination. Model is incorrect – if season and year are blocks, then season*year must be included. The authors state that year and season are random effects – this is not necessary is they are serving as blocking. In any case, if these were really random effects, then it is likely that MSE is NOT the appropriate denominator for the F-tests. Contrary to the author's assertion it is NOT a split-plot design – rather it is a variant of an incomplete block design.

In BACI designs, the interaction term is the prime term of interest – it indicated if the difference between upstream and downstream changed from before to after the impact occurred. The authors used a multiple comparison procedure if interaction was detected – but again, there is only one contrast of interest and this is of interest regardless if interaction was detected.

4.3.2.2 Results and Discussion

Unclear if the PCs were constructed using pooled data if they are plotting different definitions of PCs?

Table 4.22 – were the analyses done on the log(concentration) or the raw concentrations? Again, sample size indicates that pseudo-replication occurred.

Table 4.24 should estimate the change in the difference between upstream and downstream (with a standard error) for ALL comparisons as this is the real story.

o) Section 4.3.3 – Ability to detect change

Power analyses likely wrong because of pseudo-replication. I didn't see a power table.

Comments about power analysis in the case of interaction are not correct – there is only contrast of interest in the Muskeg River comparison so a power analysis easily done. The authors have misinterpret the intent of Steidhl (1997) – they have problems with retrospective power analyses if you use these to explain why your particular test didn't work – there is no problem in using the results of an existing experiment to predict future power. As well, the author should take Steidhl (1997) to heart and produce far more point estimates and confidence intervals.

4.3.3.2 Spatial Trends

The report computes the minimum detectable difference for a specified power than an "effect size". However, the report treats observations within a season/year combination as the indepedendent replicates when, as noted earlier, these are pseudo-replicates. The "n" in the power analysis refers to the number of blocks, i.e. the number of year/season combination as this is the "experimental" unit in question. All results are incorrect in this section.

Table 4.25 legend talks about "abundance" data which are not discussed in this chapter.

The author are surprised that for one variable the observed difference was less than the minimum detectable difference but was no declared statistically significant. However, even with an 80% power, there is still a 20% chance that a difference of that magnitude will not be detected – perhaps the study was just "unlucky".

p) Section 4.3.3.3 Conclusions and Recommendations

The report recommends that baseline data be expanded from 3 to 5 years of data but this report does not provide evidence to back up this assertion.

q) Section 4.4.3 Idenfying changes related to human activity

As the report indicates a BACI design is the minimum requirement to detect changes in environmental impact studies.

r) Section 4.5 Conclusions

It is unclear how the PCs will be used in the future as these will change depending upon the data collected – i.e. more samples are used and the component weights will change as more data are added. Consequently, estimates of means and standard deviations of current PC are not very informative.

s) Section 4.5.2.3 Ability to Detect Change

While I agree that a longer baseline is useful, this report provides insufficient justification for moving from 3 to 5 years. Unfortunately, in my experience, I suspect that 5 years will be insufficient to detect important biologically important difference! This aspect of the report needs to be reworked and strengthened.

3.4 Chapter 5 comments

The report is unclear on exactly how much sampling is done for sediment. For example, Table 5.1 appears to show that for the Athabasca river, that a single sample was taken in 2001 on the west bank upstream of Donald Creek. Unfortunately, taking a single sample at each location/bank combination provides no information about the variation at each site within a year, and unless strong assumptions are made about interactions (e.g. no year by location interaction), statistical tests cannot be performed. This needs to be clarified and duplicate sample should be taken at some (preferably all) bank/location/year combinations. These samples should be far enough apart so that they provide useful information on the variation within a year/location/bank combination, otherwise it is implicitly assumed that there is NO variation within a year/location/bank combination.

t) Section 5.1.1

With so few sites sampled over time, detecting changes over time will be difficult. Some consideration should be given to implementing a panel design.

u) Section 5.2.1 Methods

See earlier comments about eliminating 0 values. Authors have misinterpreted Zar (1984) about using the arcsin transformation. This is to be used ONLY for count data that has been expressed as a percentage – not for compositional data such as derived from this analysis of silt samples.

v) Section 5.3.1.1 Methods

Why were the modifications for Sen's method used here and not previously?

w) Section 5.3.2.1 Methods

Again, the lack of proper experimental design/data makes modeling difficult. As noted earlier, some replicate samples at the same location/year/bank need to be taken to obtain an estimate of local variation without making strong assumptions. For example, without replicate samples, it is necessary to assume that there is NO variation within a particular location/bank/year combination among replicate samples.

It is not necessary to drop terms from the ANOVA model if they are not statistically significant and the original model can be used to extract all the relevant information. This follows the principle that a non-statistically significant result does not necessarily mean the non-existence of an effect.

Rather than doing multiple comparisons looking for all possible differences among the pairs of yea/locations/bank combinations, focus in on interesting comparisons – typically among locations only.

x) Section 5.3.2.2 Results

Figures 5.13 and 5.16 look strange. The two PCs are supposed to be orthogonal to each other by the method of construction, yet the plots appear t show a distinct relationship between the two components?

5.3.3.1 Temporal trends

The requirement to expand sampling to 6 years only looks at the minimum technical requirement – it doesn't consider the actual size of the trend. While the Mann-Kendall test is "non-parametric" and only uses the relative magnitudes of the data points, its performance does indeed depend upon the actual slope of the line and the residual variation. For example, a very strong slope with small variation would imply that a monotone pattern in the points would occur often, while the same slope with a large residual variation would be less likely to have a monotone trend. An example of this is seen in Figure 5.20 of the report. Consequently, a proper power analysis would examine various combinations of effects, for example what is the chance of detecting an average 10% decline over 5 years under the variation seen in the data collected so far. The recommendation in the report that six years of data need to be collected is too simplistic.

The recommendation of accelerated sampling is pseudo-replication and is not recommended. As the report indicates, taking all samples within a single year makes no sense.

y) 5.3.3.2 Spatial Trends

Treating the east/west bank as replicates is likely fine for examining location differences within a particular year, but cannot serve as replicates for differences among years. Refer to initial discussion about experimental design for some of the perils of this recommendation.

z) 5.3.3.2 Spatial Trends – Results and Discussion

The discussion of the relative sources of variation is confusing because the ANOVA model that the report used lumps all variation into one term. There are several sources of variation – not all of which are important for detecting each type of difference. The comments about increasing sample size leading to decreased in error term in the ANOVA are wrong – increasing effort does not lead to a reduction in the various components of variance – it does lead to improved precision.

The recommendation about increasing sampling effort in to detect difference at Donald Creek need careful review to ensure that the proposed sampling design match the principles of good experimental design as outlined in the introduction.

Section 5.5.3.3.
Recommendations on increasing sampling effort in baseline are simplistic and based only on minimum technical requirements to do the computations – a proper power analysis needs to be done.

3.5 Chapter 6 Comments:

Much of the conclusions in this section are limited by the small number of years of data collected (usually 2 or fewer). The report also makes a very valid point that without reference streams that are not subject to impact, it is impossible to separate temporal effects from impact effects.

Section 6.1.1 (page 6-25)
"Individual samples collected from the same site do not represent replicates in the statistical sense because they are not independent. Widely-spaced samples from a reach (each sample representing a site) were used as replicates to compare reaches." Exactly the point made in the introduction to this report. The proposed design as listed on the bottom of page 6-5 is exactly the type of description of sampling plans needed in the other chapters of the report.

Section 6.1.3.3.
"Sampling designs have changed over time; for example, historical data and 1998 RAMP data were collected at individual sites with closely spaced replicate samples, whereas subsequent RAMP surveys concentrated on several km long reaches, with single replicates at each site." This again illustrates the need for very well documented data files so that later researchers can see exactly how survey were conducted.

(1) Section 6.2.1.1

The Principal Component analysis was done on the covariance matrix rather than the correlation matrix. Unfortunately, using the covariance matrix implies that the principal components are highly influenced by very abundant species as these are often the most variable as well. In most cases when the covariance matrix is used, the first principal component will simply be related to total abundance and is not very informative, i.e. in sites where there are lots of invertebrates, all taxa have higher abundances compared to places where abundances are lower. As well, basing results on the covariance matrix implies the results are not unit-independent, i.e. expressing densities on a different scale could change the results. A PCA on the correlation matrix is recommended.

aa) Section 6.2.1.2

As expected because the covariance matrix was used in the PCA analysis, the first component is essentially total abundance. The second component measures contrasts among three taxa.

bb) Section 6.3.1.3 – Appropriateness of study design

Report is quite correct that some reference rivers are needed in order to separate temporal trend from environmental impact trends.

cc) Section 6.3.1.4 – Conclusions and recommendations

The assertion by the authors that detecting temporal trends requires sampling a fixed locations is not correct. It is true that fixed monitoring stations often have a greater power to detect temporal changes, but sampling designs with new stations at each time point can also detect changes.

dd) Section 6.3.2.4 – Appropriateness of study design

“Representativeness” is induced by random sampling. Just because the distibution of a species is patchy, does not imply that a single sample is not “representative”. I suspect that the authors meant that small sample sizes imply that results are extremely imprecise, i.e. have a large confidence interval.

The recommendation to reduce sampling costs by reducing the number of sites measured but increasing the sub-samples per site needs further investigation. In particular, some sub-sampling should be done to see the relative sizes of the within-site and among-site variations so that an “optimal” allocation of effort across sites and within sites can be determined. The current data does not provide sufficient information to make this assessment.

ee) Section 6.3.2.5 – Conclusions and Recommendations

In general I concur with the suggestions, but some additional data needs to be collected

before committing to a long term change in the number of sites measured. What is needed is some sub-sampling at the sites to establish the within-site and among-site ratio of variation.

3.6 Chapter 7 comments

Nothing much can be done with this part of the study because of the many one-off studies conducted over the years. There is a need to standardize what will be done over the next few years.

ff) Section 7.2.1.1.- Methods

The authors compared length-frequency distribution using a repeated measures design and a non-parametric method (Page 7-13) based on classifying the data into length classes. A more direct and more appropriate analysis would use the raw data and to compare the cumulative frequency distributions using a Kolmogorov-Smirnov test or to use the binned data and use a log-linear model or chi-square test.

Figure 7.8 and similar figures. Plot the log(weight) on the Y axis; show all the data with the fitted lines for each year/season as needed. See for example, Figure 7.44 which is close, but it would nice to see the data points as well.

No mention was made of the formal analysis of fish health – but this is straight forward and is easily done using logistic regression (for percent of abnormalities) or ANOVA for the external pathology index

4. SUMMARY RECOMMENDATIONS

1. **Ensure adequate replication**. All future monitoring plans should be reviewed to ensure that real replicates will be available so that the proper statistical comparisons can be made with a minimum of untestable assumptions.

2. **Review existing analyses for pseudo-replication.** Existing analyses should be reviewed to ensure that pseudo-replicates have not been used in place of the real replicates. This will impact the reported power analyses as well.

3. **Match analysis with design.** Existing analyses should be reviewed to ensure that the model used is appropriate for the statistical design. When future studies are proposed, a "mock" analysis plan should be provided to ensure that correct model will be used in the analysis.

4. **Improve reporting of results.** Decrease the use of hypothesis testing and increase the use of confidence intervals in reporting results. As part of the report, the results should be placed in context of biologically important effects. For example, graphs similar to Figure 6 found in Steidhl et al (1977) would be very useful in interpreting the results of the report:

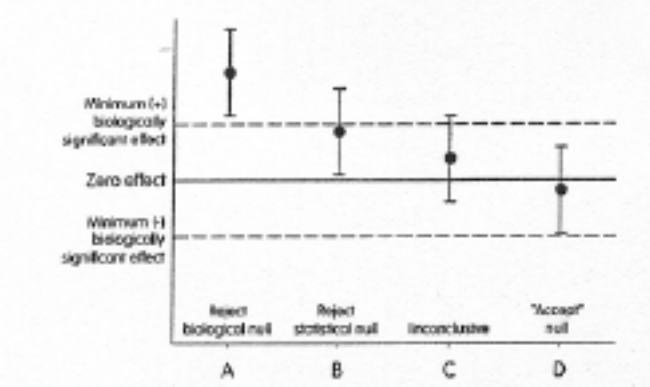


Fig. 6. Hypothetical observed effects (circles) and their associated $100(1 - \alpha)\%$ confidence intervals. The solid line represents zero effect and dashed lines represent minimum biologically significant effects. In case A, the confidence interval for the estimated effect excludes zero effect and includes only biologically significant effects, so the study is both statistically and biologically significant. In case B, the confidence interval excludes zero effect, so the study is statistically significant; however, the confidence interval also includes values below those thought to be biologically significant, so the study is inconclusive biologically. In case C, the confidence interval includes zero effect and biologically significant effects, so the study is both statistically and biologically inconclusive. In case D, the confidence interval includes zero effect but excludes all effects considered biologically significant, so the "practical" null hypothesis of no biologically significant effect can be accepted with $100(1 - \alpha)\%$ confidence.

5. **Data handing issues**. In any long term study, data storage and availability is a crucial issue. This is particularly true if contractors change during a project. Data should be available in electronic format to all interested participants. RAMP should consider setting up a separate long-term data storage/management facility whose duties would be to serve as data manager, archiver, and provider. Ideally, data could be served to interested parties using a WWW server. For example, a university could serve as a contractor. This was partially discussed in the RAMP Program Design document in the supplementary material.

6. **Consider panel designs for ongoing monitoring.** These design combine the best features of fixed and random monitoring stations.

5. REFERENCES:

Burnham, K. P. and Anderson, D. R. (2002). Model selection and inference: a practical information-theoretic approach. New York: Springer Verlag.

Cherry, S. (1998) Statistical tests in publication of the Wildlife Society Wildlife Society Bulletin, 26, 947-954.

Hurlbert, S. H. (1984). Pseudo-replication and the design of ecological field experiments. Ecological Monographs 54, 187-211.

Johnson, D. H. (1999) The Insignificance of Statistical Significance Testing. Journal of Wildlife Management, 63, 763-772.

Morrison, M. L., Block, W. M., Strickland, M .D. and Kendall, W. L. (2001). Wildlife study designs. Springer: New York.

Steidl, R. J., Hayes, J. P., and Shauber, E. (1997). Statistical power analysis in wildlife research. Journal of Wildlife Management 61, 270-279.

Weins, J. A., and Parker, K. R. (1995). Analyzing the effects of accidental environmental impacts: approaches and assumptions. Ecological Applications, 5, 1069-1083.

APPENDIX A TO BIOSTATISTICS REPORT
COMMENTS ABOUT THE INTERIM REPORTS

A less detailed review of the interim reports was conducted as many of the reports are simple data summaries with the multi-year analysis deferred to the Five-Year report reviewed above. The 1999 report has many good features in its use of statistics. It could serve as a model for all the other reports.

A.1 1997 Report

Some real replication occurred:
There appears to be some replicate sampling for some the data collected. For example, in Water Quality testing, multiple samples were taken on either bank etc. From the data in Table 3.1, the variation among samples within the same site seems to nearly as large as the variation among sites. This highlights the importance of real replication during the sampling protocols. The information from this Table may be useful in determining how many replicate samples will be needed for future studies.

Incorrect distinction between standard error and standard deviation:
Section 3.47 Standard Error and Standard Deviation

> "Standard error (SE) and standard dev**i**ation (SD) both express the var**i**ab**il**ity of resu**l**ts around the mean. However, standard error takes the sample size into consideration when calculated. By including sample size, SE gives an indication of how well we've measured the entire population. This is particularly true if you have very different sample sizes for the groups you are comparing; the larger the sample size, the more confidence you have that the data represents the popu**l**at**i**on. Standard error is calculated as: S**E**=SD ÷ 5/11;where n=sample s**i**ze. M**i**crosoft Exce**l** will calcu**l**ate SD automatically. In order to calculate SE the formula in Excel would be " =StDev(cells with data)/(sample size)^0.5 ". The "A.05" denotes square root (by asking excel to calculate to the power of 0.5). Standard error **i**s now considered to be the appropriate measure to use in any technical presentation of data and should be used in any figures or tables of fish population statistics."

This is mostly incorrect. Standard deviations measure the variation of individual measurements around the mean. Standard errors measure the precision of an estimate. [Technically, the standard error of an estimate measures the variation of the estimate if repeated samples of the same size were taken from the population.]

The formula quoted above is ONLY valid for the standard error of a mean collected under a simple random sample. It is NOT valid for other estimates nor for other designs. However, it is possible to compute a standard error for other estimates and for other designs.

Standard deviation should be used when variation of individuals values is to be highlighted. Standard error is to be used when the estimate of the underlying population parameter (e.g. the population mean) is to be highlighted.

A.2 1998 Report

gg) Section 3.1.1.2 Field Methods for water and sediment quality – some real replication?

Figure 3-1 indicates that for the most part there was no real replicate sampling, but at the very top site (on the map), there may be real replication.

hh) Section 3.2.2 Benthic Invertebrates field methods – some real replication?

Some replication done here?

Section 4.2, Table 4-8 – illustration of dangers of p-values.
This is an illustration of one of the "dangers" of p-values. The p-values reported here are for a test that the slope of the size at age curve is zero. Yet, this is a silly hypothesis because it known to be false and not biologically interesting. It would be much more informative to report confidence intervals for the slopes and intercepts.

Similarly, in the discussion of the comparison of results between 1997 and 1998 (page 4-11), it was stated that the intercepts were "significantly different", yet no value was given for the estimated difference along with a standard error. If the estimated difference was .001 with a se of .0001, who cares? Table 4-9 is much more informative and should be the standard way of presenting such results.

In graphs similar to Figure 4-5, the Y axis should be in relative frequency (e.g. %) rather than absolute frequency.

Figure 4-12, please show the raw data as well so that it can be seen if the observed change in the regression line is "caused" by a few anomalous fish.

A.3 1999 Report

ii) Section 3.1.2.2 – Good practice for multiple comparisons

"To control experiment-wise error, a significance level of p=0.017 (i.e., £\/no. of comparisons, 0.05/3) rather than p=0.05 was used (i.e., Bonferroni's adjustment)."

This is a good practise that should be extended to the other reports. This report also has a good discussion of power analysis and biologically meaningful difference.

This report presented statistical issues well and did power analysis well. It should serve as a model for future reports.

A.4 2000 Report

jj) Table 4.7 – some information on real replicates variation?

Table 4.7. Here is some information on local variation of sediment values among replicate stations. I disagree with the conclusion that the variation is small – it looks rather alarmingly large often varying ±50% of the mean value! This information should be used to establish the number of replicates needed at individual sites for future sampling plans.

A.5 2001 Report v. 1

kk) Misunderstanding about the use of the arcsine transformation

Page 3-52. Misunderstanding about the use of the arsine transformation – not necessary for compositional data such as LSI or GSI. The arcsine transformation is only appropriate for proportions that are derived from counts of discrete objects, e.g. what proportion of fish have lesions, where the binomial distribution is the underlying description of the data. Compositional data does not follow a binomial distribution and so the use of this transformation is inappropriate.

Misunderstanding of the Kruskal-Wallis test.

Page 3-65.

> "The Kruskal-Wallis test is used instead of Analysis of Variance when samples do not come from normal populations, are heterogeneous or do not have equal numbers of data in each group (Zar 1999). The Kruskal-Wallis test was the appropriate test to use for RAMP since different lakes were sampled using varying numbers of transects and plots. The test was applied to identify significant differences between the lakes for vegetation groups, species, and water chemistry."

This is not correct. Non-parametric tests, despite their name, also have assumptions. For example, they assume equal variances in all groups. It is not necessary to use the KW test if sample sizes are unequal in groups, and it is not appropriate if the variances are heterogeneous. They also require the same attention to matching design and analysis, i.e. the KW test assumes a single factor completely randomized design. Designs with transects and plot within transects are NOT completely randomized designs, and consequently should not be analyzed using a KW test, nor with a single-factor CRD ANOVA.

A.6 Ramp Program Design and Rationale1

Section 4.2 Sediment sampling – compositing vs real replicates.

Page 4-5 .

> "At each sample site, except upstream of Fort McMurray and upstream of the Embarras River, one composite sample will be prepared every fall by combining 4 to 6 grab samples collected from depositional areas located between the east river bank and 25% of the river width (Table 4-1). The process will then be repeated between the west river bank and 25% of the river width."

The compositing is fine as the replicates samples within a small physical area in the sample size are pseudo-replicates, but there is no real replication at the sample locations. I would also take some real replicates (see my earlier report) at some sites, i.e. move 200 m upstream or downstream, to identify the actual within site variability.

Need for real replication:
Table 4-3 needs to be expanded to indicate how many real replicates will be gathered – at the moment, only a single replicate is gathered at each sampling location.

ll) Reallocation of resources from split/duplicate samples?

In the QA/QC the program is willing to spend some money on split-samples. Perhaps divert some of this money to real replication.

How was it known that three samples composited will be enough?

mm) Section 5.1.2 Benthic sampling – real replication used, but better rationale needed

Here the necessity for real replication is explained. Why were 15 samples taken – what is the rationale for this?

nn) Section 6.1.1.2 Fish Inventory – Dangers of CPUE as adundance measure

> "Species distribution, composition and relative abundance (i.e., catch per uniteffort) will be recorded."

CPUE to measure abundance is notorious poor because of changing gear, changes in catch efficiency over time, difficulty in standardizing etc. I suspect a better measure of impact would be fish health indices, composition (young vs old), and length-frequency shifts.

oo) Section 6.1.2.2 Mackay River Fish Inventory – more rationale needed about tagging

Fish are to be tagged, but sampling will be done very three years? Who will return tags that hare added? Will the tags last three years?

pp) Section 6.3.1.1. Muskeg River radiotelemetry – unclear purpose?

What information will be gained by this monitoring?

Section 6.3.4 Database development for RAMP Fisheries – need for implementation
This is a major issue for all components of the RAMP program – how will data be stored, accessed, and protected during the life of the monitoring program? This is covered in Appendix 1 with a power point presentation on the FWIS – is this available to RAMP?

Section 7 – Vegetation surveys –cluster/two stage sampling
Many of the vegetation surveys have implemented real replication, but subsequent analyzes need to take into account the cluster/two-stage sampling design:

"In 2001, 11 plots were located on 6-transects with approximately two plots per transect." (Shipyard Lake study).

Section 7.1.1.5 Reference wetlands – panel design should be considered?
Good that at least two reference wetlands are being measured. The report mentions the possibility of bringing in new lakes if problem sarise with the control lake – try and get as much advance notice as possible of this. Perhaps plan for a panel design from the start.

Section 8 – Gradient exposure designs.
Sampling plan is appropriate with a gradient in exposure and spatial controls that will not be exposed. If the oil sands expands, will the gradient in exposure change over time? This was an issue for the TEEM monitoring project where the expansion of the oilsands has exposed many of the "control" sites to deposition.

Appendix IV. Oil Sands Regional Aquatic Monitoring Program (RAMP): Scientific Peer Review of the Five Year Report (1997-2001): Reviews of Individual RAMP Components.

TABLE OF CONTENTS OF APPENDIX IV

For the convenience of readers Appendix IV has been separated into seven electronic files: one for each component. The files are named "2004 RAMP review- hydrology template.doc, 2004 RAMP review-sediment quality template.doc... etc.". Each electronic file includes the Introduction and template description sections followed by the component report following the prescribed template.

INTRODUCTION	1
TEMPLATE FOR REVIEW OF COMPONENTS AND INSTRUCTIONS TO REVIEWERS	2
CLIMATE AND HYDROLOGY	CH7
Characterizing Existing Variability	CH7
Detecting Regional Trends and Cumulative Effects	CH15
Monitoring to Verify EIA Predictions	CH25
WATER QUALITY	WQ7
Characterizing Existing Variability	WQ10
Detecting Regional Trends and Cumulative Effects	WQ16
Monitoring to Verify EIA Predictions	WQ20
SEDIMENT	S7
Characterizing Existing Variability	S11
Detecting Regional Trends and Cumulative Effects	S16
Monitoring to Verify EIA Predictions	S25
BENTHIC INVERTEBRATES	BI7
Characterizing Existing Variability	BI12
Detecting Regional Trends and Cumulative Effects	BI16
Monitoring to Verify EIA Predictions	BI21
FISH POPULATIONS	FP7
Characterizing Existing Variability and Detecting Regional Trends and Cumulative Effects	FP7
Monitoring to Verify EIA Predictions	FP16
Fish Abnormalities	FP18
AQUATIC VEGETATION	AV7
ACID SENSITIVE LAKES	ASL6