2010 REGIONAL AQUATICS MONITORING PROGRAM (RAMP) SCIENTIFIC REVIEW

RAMP Review Panel:

Dr. Burn
Dr. Dixon
Dr. Dubé
Dr. Flotemersch
Dr. Franzin
Dr. Gibson
Dr. Munkittrick
Dr. Post
Dr. Watmough

Submitted By:

Catherine Main, M.Sc., MCIP, RPP, P.Geol., P.Geo.
Program Leader
Integrated Water Management Program
Alberta Innovates – Technology Futures
3608 – 33 Street N.W.
Calgary, AB T2L 2A6
Tel. 403-210-5369
Catherine.Main@albertainnovates.ca

January 6, 2011

Table Of Contents

List of Appendices

1.0 INTRODUCTION

1.1 Background

The Regional Aquatics Monitoring Program (RAMP) is a multi-disciplinary initiative established in 1997 to determine, evaluate and communicate the state of the aquatic environment and any changes that may result from cumulative resource development within the Regional Municipality of Wood Buffalo of northern Alberta (RAMP, 2009). The Program seeks to monitor changes in hydrology, water quality, benthic invertebrate communities, sediment quality, fish populations and acid sensitive lakes.

1.2 Review Approach

The purpose of the 2010 RAMP review is to evaluate the methods presently used by RAMP to evaluate aquatic ecosystems and suggest changes to update the existing program where warranted. The program strives to achieve a holistic understanding of potential effects of developments in the Athabasca Oil Sands Region so that long-term trends can be identified, cumulative effects can be assessed, and potential impacts can be addressed.

The overall goal of the RAMP program review is to answer three key questions as they pertain to each of the aforementioned areas of concern:

1. Can the present Program detect changes if they occur?

2. Can the source of any potential changes be identified by the present Program?

3. Are the appropriate questions being asked by the Program and are the appropriate criteria being monitored to answer those questions?

A review team was selected using a documented selection process. The RAMP Review Panel is composed of individuals with specialization within the following areas:

- Climate and Hydrology: Dr. Burn and Dr. Gibson
- Water Quality: Dr. Dixon and Dr. Dubé
- Benthos and Sediment: Dr. Munkittrick and Dr. Flotemersch
- Fish Populations: Dr. Post and Dr. Franzin
- Acid-Sensitive Lakes: Dr. Gibson, Dr. Dixon, Dr. Watmough

Reviewers were selected with the intention of having a minimum of two reviewers in each area of specialization. The reviewers were asked to conduct a thorough review

within their areas of specialization evaluating whether the program met its objectives specific to their area(s) of expertise. In addition, each reviewer was asked to comment on the program as a whole.

Information provided as resources for the review included but was not limited to:

- Present Monitoring Work Plan (Applicable to 2012)
- 2009 Technical Report (RAMP, 2009) and the previous reports
- 2009 Technical Design and Rationale Document (RAMP, 2009)
- 2004 RAMP Scientific Peer Review Report (Ayles et al., 2004)
- RAMP Terms of Reference (RAMP, 2009)
- An Assessment of the Regional Aquatic Monitoring Program (RAMP) Fish Survey (Whittier and Hughes, 2008)

The following report presents a synthesis of their recommendations. First, RAMP's present ability to successfully address three key questions posed as part of the goal of the RAMP review (as outlined in section 1.2 of this report) and RAMP's present ability to achieve the program's objectives as stated in the terms of reference for RAMP were evaluated in section 2.0. The report outlines the results of each individual review for each component in section 3.0. Each individual review is appended (Appendices A to I) at the end of the report. Some of the reviews found within the appendices have addendums attached to them. The addendums were attached following discussions between reviewers with the purpose of clarifying priorities and key issues as outlined by the individual reviewers. Recommendations for the scientific implementation and the program management of the RAMP program are presented (section 4.0). A paragraph outlining the background of each reviewer is included in section 5.0.

2.0 REVIEW OF RAMP OBJECTIVES

The RAMP review was evaluated based on its ability to meet the RAMP review goal (section 1.2) and the RAMP program objectives. The reviewers believe that the existing RAMP program does not successfully address the three key questions posed in section 1.2.: the present program is not sufficient to detect changes if they occur; the present program cannot sufficiently identify potential sources resulting in the change(s) if changes are detected; and not all of the appropriate questions are being asked by the RAMP program and appropriate criteria being monitored to answer those questions.

The RAMP program objectives, as outlined in the RAMP Technical Rationale Document (RAMP, 2009), are listed below. The RAMP Review Panel evaluated whether RAMP is successful in meeting its program objectives and reasons for success or failure.

1. ***Monitor aquatic environments in the oil sands region to detect and assess cumulative effects and regional trends;***

The RAMP program has not met this objective. The program is currently incapable of detecting regional trends and cumulative effects because of the program design, the loss of reference sites, inadequate representation of variability (inflated variability, Type II error), the potential alteration and contamination of reference sites, and lack of integration with other monitoring in the basin that would provide more spatial and temporal coverage.

2. ***Collect baseline data to characterize variability in the oil sands area;***

The RAMP program has not met this objective. The fact that baseline sites are identified in the program is a positive aspect, although the integrity of those sites needs to be examined. There is inadequate spatial and temporal coverage of baseline data within the program to adequately assess the baseline variability. Secondly, the existing reference sites may not adequately characterize baseline variability based on the reasons listed under objective 1.

It is recognized that the existing scale of the monitoring program makes it difficult to recognize or identify regional baseline sites that may not already be compromised or impacted by anthropogenic activities. There should be more integration with the airshed and groundwater monitoring programs to accurately assess and characterize the baseline sites.

3. ***Collect and compare data against which predictions contained in Environmental Impact Assessments (EIAs) can be assessed;***

The RAMP program has not met this objective. The 2004 RAMP Review (RAMP, 2004) recommended that a summary of impact predictions for the EIAs be compiled and that RAMP test the predictions of the EIAs. A summary of the indicators designated within EIAs were included as part of the 2009 Technical Design and Rationale Document (Table 2.12 and Appendix 3), however an evaluation of the accuracy of the EIA predictions has not been completed. Inconsistencies in the indicators and impact criteria between the existing EIAs and the RAMP monitoring program make it unclear if the predictions can be validated. There should be an assessment of which predictions have the potential to be validated with the existing RAMP monitoring program.

4. ***Collect data that satisfies the monitoring required by regulatory approvals of oil sands developments;***

Based on reports and information provided as part of the RAMP program, it cannot be determined whether the monitoring requirements of approvals for each individual oil sands development have been fulfilled. The RAMP Review Panel recommends that the compliance monitoring be integrated into a broader monitoring strategy that includes RAMP.

5. ***Collect data that satisfies the monitoring requirements of company-specific community agreements with associated funding;***

Based on reports and information provided as part of the RAMP program, it cannot be determined whether the monitoring requirements of company-specific community agreements have been fulfilled. The RAMP Review Panel recommends that the monitoring be integrated into a broader monitoring strategy that includes RAMP.

6. ***Recognize and incorporate traditional knowledge into monitoring and assessment activities;***

No information regarding Traditional Ecological Knowledge (TEK) was provided for review as part of the RAMP program; however, the RAMP Review Panel recommends that TEK be integrated into a broader monitoring strategy that includes RAMP.

7. ***Communicate monitoring and assessment activities, results and recommendations to communities in the Regional Municipality of Wood Buffalo, regulatory agencies and other interested parties;***

An improved communications strategy for the release of data and reports is required.

8. ***Continuously review and adjust the program to incorporate monitoring results, technological advances and community concerns, and new or changed approval conditions;***

The RAMP program has partially met this objective. The RAMP Review Panel recommends the creation of an external Science Advisory Panel to provide continuous, hands-on oversight. An External Science Advisory Panel is necessary because the RAMP program requires continual review and adaptation relative to the recent literature, advances in science and external experience. This external panel should work concurrently with the RAMP Technical Committee.

9. ***Conduct a periodic peer review of the program's objectives against its results, and recommend adjustments necessary for the program's success.***

The RAMP program has met this objective. However, the RAMP Review Panel recommends that the 5-year RAMP Scientific Peer Review should be continued using a review panel composed of experts that are separate from of the External Science Advisory Panel identified under objective 8. The review process should ensure that the integration across components is addressed before delivery of a final report.

3.0 COMPONENT REVIEW RESULTS

In answering the three key questions posed as part of this RAMP program review goal (section 1.2), the reviewers have highlighted some of the limitations in the current program and have identified a number of modifications that would significantly improve RAMP's ability to monitor and to identify changes in the aquatic environment.

3.1 Climate and Hydrology Component

The climate and hydrology component of the RAMP study was conducted as a main part of the reviews of Dr. Burn and Dr. Gibson. Individual review submissions are appended in Appendices A and B respectively. The results of the individual reviews identified the following as the main areas requiring improvement or change:

- More climate and hydrological data are required.
- Current models need to be revised to avoid too many assumptions and inaccurate estimations and predictions.
- Focus on long-term trend analysis by maintaining a high number of monitoring stations.
- Need for groundwater monitoring in the program as well as assessment of groundwater-surface water connectivity and interaction; modify water balance approach to include groundwater.
- Design of monitoring network needs to be proactive rather than reactive.

Dr. Gibson suggests the addition of climate stations, specifically in the region south of Fort McMurray, eliminate the need for using interpolated values during data interpretation. Currently, regionally distributed climate data are limited, and data that are available are not always available through the online data access point. The availability of sufficient climate data is important for hydrological modelling, although both reviewers indicate limitations of the current water balance model must also be addressed as estimates garnered from model runs may inaccurately represent baseline conditions and change source. Some of assumptions used within the models do not account for hydrologic changes expected in the study area. These include changes in seasonal or inter-annual variability in soil moisture and runoff, catchment

responsiveness, storage, groundwater /surface water interactions, lake levels, and wetlands. Significant expansion in lake level monitoring is recommended. Where possible, these activities should be conducted at sites sampled as part of the acid-sensitive lakes monitoring component of RAMP.

The addition of groundwater monitoring to the program will help improve the understanding of the water balance of the region. Groundwater/surface water interactions have a strong influence on discharge, water quality and wetlands. An understanding of the role of groundwater will be necessary to predict the hydrological changes that can be associated with increased development.

Dr. Burn indicated that the monitoring network needs to be revised based on a long-term vision of anticipated oil sands development rather than a reactive approach based on development as it occurs. This network must consider baseline as well as test site requirements. The network design process needs to anticipate the development of oil sands properties and locate gauging stations both upstream and downstream of potential development sites to ensure that:

1. Baseline conditions are continually monitored for the undisturbed portion of the watershed; and

2. Downstream baseline records in the watershed, for the period prior to development, are sufficiently lengthy to form a strong basis of comparison with test conditions measured after the development of oil sands.

Consideration should be given to the concept of developing a "regional" network of gauging stations, consisting of stations in the study area and stations close to the study area. The latter should be stations that can be considered to be hydrologically similar to the stations within the study area.

3.2 Water Quality Component

The water quality component of the RAMP study was conducted as a main part of the reviews of Dr. Dixon, and Dr. Dubé. Individual review submissions are appended in Appendices C and D respectively. The results of the individual reviews identified the following as the main areas requiring improvement or change:

- Increase monitoring of lakes and other surface water features, especially baseline sites.
- Include naphthenic acids (NAs) and polycyclic aromatic hydrocarbons (PAHs) as part of the monitoring program in the water column as well as the sediment pore

water. Consideration should be given for the use of petroleomics in water analysis.

- Consider studies on seasonal variability in water quality in relation to results obtained from the annual fall sampling program.
- Develop a standard protocol for changes in analytical procedures over time.
- Justify rationale for the river mouth sampling stations use as the watershed "cumulative effects" indicator stations.
- Create consistency of impact criteria used in Environmental Impact Assessments (EIAs) and in RAMP
- Clarify the method and application for use of the Water Quality Index.

The reviewers indicated a need for new surface water body monitoring stations (supported by Dr. Gibson). These new stations would be selected based on a need to maintain an appropriate ratio of test to baseline sites. Dr. Dixon suggests that although monitoring of water quality for acid sensitive lakes (ASLs) is occurring at sufficient levels to meet objective 2 of the program, monitoring designed for biological components are being met only at two test and two reference sites for lakes.

The reviewers indicate that it is important to measure NA, PAHs, and metals as part of the monitoring program to assess potential contamination of reference sites by atmospheric deposition. A more thorough characterization of the organic compounds using techniques such as Petroleomics (Fourier transform- ion cyclotron resonance mass spectrometry, FT-IRCMS) was proposed as a compliment to NA partitioning.

Dr. Dixon recommended the inclusion of temporal monitoring to the current program design. Variability in seasonal events may also create variability in contaminant transport, thus necessitating some seasonal monitoring to assess the most vulnerable times and locations.

Since changes in analytical procedures will undoubtedly occur over time, Dr. Dixon suggested that a standard method should be developed to transition from one sampling analysis procedure to another so that the data can be compared over time.

Dr. Dubé emphasized the need for clarification of test site locations with respect to developments and land disturbances. Understanding the local and regional variability between baseline sites will establish the justification for future baseline station selection. Dr. Dubé suggests this may be accomplished by creating an overlay of new land disturbances, point source discharges, and outcrops of the McMurray Formation that can be presented with the locations of monitoring sites. In the absence of understanding where monitoring stations are relative to development activities, within the geologic

framework, and within the watershed, justification is needed for blanket use of river mouth sampling sites to assess cumulative effects.

Dr. Dubé and Dixon both indicated a need for consistency in impact criteria used in the EIAs and the RAMP program. Furthermore, Dr. Dubé indicated that the impact criteria must be tied to some level of decision or action in future EIAs and for RAMP.

Dr Dubé indicated that the Water Quality Index can not be compared spatially with different parameters and benchmarks. Clarification is required regarding the method and application used.

3.3 Benthos and Sediment Quality Component

The benthos and sediment quality component of the RAMP study was conducted as a main part of the reviews of Dr. Munkittrick and Dr. Flotemersch. Individual review submissions are appended in Appendices E and F respectively. The results of the individual reviews identified the following as the main areas requiring improvement or change:

- Increase in sampling along the mainstem Athabasca River.
- Sampling location design to concentrate on a localized habitat and/or potential impact location approach rather than a riffle to riffle approach.
- Sample analysis methods including statistical methods need to be revamped to reduce noise and variability in data and allow for identification of impact.
- Harmonization and integration of both RAMP components and study results outside of the RAMP program.
- Development of a model to provide benchmark or baseline for test site evaluation.

The Athabasca River is the main downstream receptor and therefore requires increased sampling effort. If impacts and effects of impact are detected in this system, the impact has already reached significant levels.

Benthic sampling needs to proceed on a multi-habitat or Functional Process Zones rather than on a riffle to riffle approach. The variability in samples from riffle to riffle is too high. Localized changes need to be first identified and detected prior to identifying wide-reaching impacts and effects and this can be done by looking in areas you would expect to potentially see impact and then evaluate if there is an ecologically relevant change and then look on a larger scale to see the extent of impact.

Dr. Munkittrick's review strongly emphasized the need to use appropriate statistical methods to both decrease variability in sampling results and the associated noise in the data. The statistical approach results in large variability and noise within the data making the identification of trends and impacts within the data difficult. Detailed suggestions for revisions to the statistical approach are given in Appendix E.

The integration or harmonization of the hydrologic, chemical and biotic components is seen as integral in the understanding of impact significance. Dr. Munkittrick and Dr. Flotemersch both have indicated the value of developing a model to improve our understanding of the relationship between hydrologic, chemical and biotic constraints for baseline sites against which test sites can be compared (mechanistic model) and to provide predictive capabilities in light of baseline site reduction.

3.4 Fish Populations Component

The fish population component of the RAMP study was conducted as a main part of the reviews of Dr. Post and Dr. Franzin. Individual review submissions are appended in Appendices G and H respectively. The fish populations section had more suggestions for revisions than any other component. The reviews identified the following as the main areas requiring improvement or change:

- Use probabilistic fish sampling methods to capture temporal and spatial variability within the watershed.
- Use physiological indicators to assess exposure.
- Increase the number of reference sites for sentinel species.
- Define reference areas in the Athabasca River mainstem.
- Use age-based demographic analysis.
- Use individual fish analysis rather than means in statistical analyses of metals in fish tissues.
- Explore alternative sampling methods to enhance capture of small-sized fish.
- Do full assemblage sampling while collecting sentinel species.
- Return to lethal sampling of sentinel species so that the gonadal somatic index endpoint, fecundity and tissue analyses can be done to maximize information for the effort.
- Include all available historical pre-RAMP data.
- Improve integration with all other components.

Dr. Franzin identified the lack of any physiological indicators that would be effective in identifying toxicology or stress in fish samples collected. It is expected that these

indicators may be able to detect potential impacts prior to those of more physical measures. Dr Franzin also identified that there were only two reference sites for sentinel fish species and that these need to be increased in number. Dr. Post indicated the need for more wide-spread random spatial sampling within the watershed. Dr. Post also outlined the need for a habitat assessment model approach to sentinel species assessment which includes the integration of physical, chemical and biological data.

Both reviewers made suggestions in terms of assessment and sampling approach. Both reviewers indicated a need to define and increase the number of reference sites on the Athabasca River mainstem. In particular, an increase of sites for monitoring of sentinel species was recommended. Both reviewers recommended that an age-structured approach was more useful to assess impact rather than simple body size assessments. In addition, tissue analysis should use individual fish results rather than population means which would provide more information on potential impact. Graphics and data analyses should be conducted on individual fish to maximize power of the analyses. In addition, Dr. Post recommended that a return to lethal sampling of sentinel species be conducted so that reproductive data of fish can be collected and evaluated. The importance of improving methods to capture small-sized fish and including the appropriateness of young-of-year analyses was discussed by both reviewers.

It was strongly argued that integration of RAMP data within existing information and integration between components within RAMP would provide a basis for a more complete assessment of potential impacts. Both reviewers indicated that there were advantages to using background historical data that pre-exist the formation of RAMP to develop an understanding of baseline conditions. In addition, the need to integrate physical, chemical and biological data was emphasized.

The use of fish fences should be critically evaluated. If fish fences cannot reliably be operated in the smaller rivers, a technique that will yield catch per unit effort (CPUE) demographic samples, relative abundances of species and tissue samples should be employed.

3.5 Acid-Sensitive Lakes Component

The acid-sensitive lakes (ASL) component of the RAMP study was conducted as a main part of the reviews of Dr. Gibson, Dr. Dixon, and Dr. Watmough. Individual review submissions are appended in Appendices B, C and I respectively. The results of the individual reviews identified the following as the main areas requiring improvement or change:

- Development of a clearly stated sampling design strategy.

- Assessment of the adequacy of the sampling event frequency
- The current ASL survey is not representative of natural variability in landscape.
- The use of the steady state critical load component needs to be reevaluated.
- Justification of analytical methods.

The reviewers commented on the lack of justification or explanation for the sampling strategy used to select the 50 lakes used in the ASL survey. The report does not outline why standard survey practices involving stratified-random sampling design were not used to ensure that the lakes selected for the ASL monitoring included representation of the natural variability, within a range of class sizes for the region. The use of lakes more sensitive to acidification for the ASL sampling may result in a negative perception of ASL impact in the study region.

There is very limited biological monitoring of lakes in the region. Samples are collected but have not been thoroughly analyzed. The reviewers see an opportunity to expand the lake monitoring to incorporate other chemical and biological parameters.

The once a year fall sampling program was noted as potentially being insufficient to capture the state of the lakes with respect to acidification. Episodic acidification associated with events such as snow melt may not be appropriately captured by the existing sampling program and needs to be evaluated.

There are a number of comments in the appended reviews regarding assumptions used in the critical load calculations (e.g. base cation concentrations, acid neutralizing capacity, and runoff). The use of critical loads should be evaluated to determine if they are appropriate for the region (e.g. wetlands, complex hydrology, etc.) and if they are calculated correctly. Currently, multiple reports need to be reviewed to determine how critical loads were calculated and what interpretations were made based on these calculations. Further details in the calculations and clarity of how data are interpreted should be presented. Identification of trends in chemistry is currently the most useful analytical method undertaken, but the chemistry trends should be assessed on an individual lake basis rather than grouped together for a composite analysis.

4.0 RECOMMENDATIONS

4.1 Science Based Recommendations for Program

1. The Program requires an expanded temporal and spatial scope. The oil sands are a large temporal and spatial development that requires a landscape and watershed scale approach to assess its potential impacts. The approach should include a probabilistic design with elements of a Before-After Control-Impact (BACI) design assessment method embedded within a stratified random probabilistic sampling program (Smith, 2002). The Program needs to move away from a reactive design to a more strategic proactive design especially in terms of developing valid baseline and test site selection within a probabilistic design.

2. A program of this scale and magnitude is beyond the resources that are currently available within RAMP. In order to have a meaningful monitoring program in this large and complex system, integration of all monitoring programs in the region is required including groundwater, air monitoring, habitat, land use, riparian, etc. These monitoring programs include but are not limited to: Wood Buffalo Environmental Association (WBEA), Cumulative Environmental Management Association (CEMA), Alberta-Pacific Forest Industries Inc. (ALPAC), Environment Canada (EC), Department of Fisheries and Oceans (DFO), Alberta Environment (AENV), Athabasca Tribal Council (ATC) and the Regional Aquatic Monitoring Program (RAMP).

3. Acid-sensitive lakes and wetlands programs should be expanded to include biological components.

4. The monitoring program needs to be embedded within a decision-making framework which includes a definition of effect for all of the components. The decision-making framework needs to clarify the effect indicators, when an effect exists, and what actions will be taken when an effect is identified. Consideration of sensitive indicators in the monitoring design is required (ranking of relative sensitivity of different indicators to ensure efficiency in measuring effects). For some of the monitoring components, this type of effects-based assessment currently is being conducted but the approach needs to consider consistency across components.

5. There is a substantial concern that baseline variability in the existing RAMP program is inflated for some components since the approach does not distinguish between spatial and temporal variability during the statistical analysis. A revised monitoring program design must consider issues associated with measuring variability including the representativeness of the baseline sites. Refer to the

comments outlined for objective 2 above. In addition, understanding of variance in data analysis must include an understanding of methodological bias and variance.

6. Integration of the physical and biological components across sampling stations is critical. The integration of components at individual sampling stations in flowing waters is required. In addition, an understanding of how changes in individual components are linked in a systems context also needs to be considered in an integrated study design to interpret the significance of ecosystem change.

4.2 Recommendations for Program Management

1. This review, as in past reviews, has identified that there are multiple monitoring programs in the region and RAMP's objectives would be best served by integrating information from these other programs, including other historical programs (e.g., AOSERP). The model for integration needs to consider the following questions:

    - Who has responsibility for what is being monitored?
    - What role does government play in monitoring?
    - Is there a need for one overall monitoring program?

2. An independent external RAMP Revision Committee of scientists with expertise in monitoring design is required immediately to modify the existing RAMP program to enable it to meet its objectives.

3. RAMP requires an independent External Science Advisory Panel to provide continuous hands-on oversight. This panel should work concurrently with the RAMP Technical Committee and the RAMP Revision Committee.

4. The Peer Review process should be retained and is separate from the External Science Advisory Panel. A Terms of Reference should be developed for selection of the RAMP Review Panel to allow for accountability and uniformity in the selection process.

5.0 REVIEW TEAM

The RAMP Review Panel was composed of the following individuals.

- Climate and Hydrology: Dr. Burn and Dr. Gibson
- Water Quality: Dr. Dixon and Dr. Dubé
- Benthos and Sediment: Dr. Munkittrick and Dr. Flotemersch
- Fish Populations: Dr. Post and Dr. Franzin

- Acid-Sensitive Lakes: Dr. Gibson, Dr. Dixon, Dr. Watmough

An outline of their background is provided in the following paragraphs.

Dr. Donald Burn, Ph.D., P.Eng., studied at the University of Waterloo and is currently a Professor in the Department of Civil and Environmental Engineering at the University of Waterloo. Dr. Burn teaches courses in hydrology, environmental systems modelling, and civil engineering systems at the undergraduate level and courses related to water management at the graduate level. Dr. Burn is a former Co-Editor of the Canadian Water Resources Journal and is a member of the Study Board for the International Joint Commission's International Upper Great Lakes Study. Dr. Burn conducts research dealing with statistical aspects of hydrology including work on regional flood frequency analysis, drought probabilities in large drainage basins, and the hydrologic implications of climatic change.

Dr. D. George Dixon, Ph.D., (B.Sc., Sir George Williams University, 1972; M. Sc., Concordia University, 1975; Ph.D., University of Guelph, 1980) is Vice-President, University Research and Professor of Biology at the University of Waterloo. He served as Dean of Science from 2001-2007. Dr. Dixon has received both the Award for Excellence in Research and the Distinguished Teaching Award from the University. He has over 30 years experience in aquatic toxicology and environmental risk assessment and management, principally but not exclusively, with respect to the environmental impacts associated with metals and mining activity. At various times during his career he has served as an advisor on metal contamination issues to Environment Canada, the Department of Fisheries and Oceans, the Department of Justice (Canada), the U. S. Environmental Protection Agency, the U. S. National Oceanographic and Atmospheric Administration, the Department of Justice (U. S.) and the World Health Organization, among others. Dr. Dixon maintains an active research program, which at present is focused on development of methods for environmental effects monitoring, methods of assessing the environmental risks associated with exposure of aquatic organisms to metal mixtures, and on the aquatic environmental effects of oil sands extraction in northern Alberta. He has supervised the research of over 70 M. Sc. and Ph. D. students and has authored or co-authored over 230 refereed journal articles. He has also developed and taught numerous courses in environmental toxicology and risk assessment.

Dr. Monique Dubé, Ph.D. is currently a Canada Research Chair in Aquatic Ecosystem Health Diagnosis at the University of Saskatchewan. Her interest is to bring "science to service" with a focus on assessing and managing the cumulative effects of multiple stressors affecting freshwaters. Monique has 20 years experience in aquatic ecotoxicology in academia, government, and in the private consulting sector. She has published over 100 works and presented at 150 public and conference forums. She is

fortunate to work with a suite of talented graduate students and with colleagues across public, academic, government, industry (pulp and paper, mining, oil sands), consulting, and global (Global Environment Monitoring Program/United Nations Environment Program) networks.

Dr. Flotemersch. Ph.D., (B.S. in Wildlife Biology, 1987, Murray State University, M.S. in Aquatic Ecology, 1992, Eastern Kentucky University, Ph.D. in Large River Ecology, 1997, Mississippi State University) is a Senior Environmental Scientist (Research Ecologist), Ecological Exposure Research Division (EERD), Chief Ecosystems Research Branch (ERB) with the USEPA. Dr. Flotemersch's research activities have been directed towards assessing the influence of environmental stressors on health and ecological integrity of riverine ecosystems. During his tenure with EPA, he has led numerous research projects in support of the development of indicators of river ecosystem condition. Currently he is serving as the indicator lead for the Office of Water's National Rivers and Streams Assessment. His current research efforts are focused on rapid techniques for the classification of riverine resources in support of multiple EPA mission-relevant tasks (e.g., characterization of ecosystem services, asset trading, healthy watersheds initiative, assessment, restoration).

Dr. William G. Franzin, Ph.D. joined the Freshwater Institute in Winnipeg as a research scientist with Fisheries and Oceans Canada in 1975 where he worked until retiring in 2008. He was an adjunct professor in the Zoology Department at the University of Manitoba until 2005 where he supervised or co-supervised 10 graduate student theses at the master's and doctoral levels. His broad fish/fisheries research interests have included fish biogeography and diversity, effects of heavy metal toxicity on wild fish populations, fish genetics, walleye stocking, instream flow issues, invasive aquatic species and species at risk. He also worked with CEMA on the Surface Water Working Group and the Instream Flow Needs Technical Task Group first as a science representative of DFO for eight years and later for two years as a contracted Technical Program Manager for IFNTTG. Dr. Franzin has authored or co-authored 45 published papers, book chapters and reports, dozens of presentations at scientific meetings and contributed to countless departmental submissions and reviews. He was president of the American Fisheries Society in 2009. Dr. Franzin continues to be active in fisheries science as a consultant; Laughing Water Arts & Science, Inc.

Dr. John Gibson, Ph.D., P.Geo. P.Geol. is a Research Professor and Program Manager Oil Sands and Mining Water Management – Alberta Innovates - Technology Futures in Victoria BC. Dr. Gibson is an AITF research professor based at the Vancouver Island Technology Park. He is a water resources specialist with expertise in application of isotope tracer techniques to evaluate surface and groundwater issues in the oil sands region, as well as considerable experience working in regional, national and

international water projects. Dr. Gibson is Past President of the International Commission on Tracers of the International Association of Hydrological Sciences, he is founder and Past Chair of the Canadian Geophysical Union Committee on Isotopic Tracers, and has been employed to coordinate research programs on behalf of AITF, Environment Canada, the International Atomic Energy Agency, and the Australian Nuclear Science and Technology Organization. He is currently co-leader of the regional water initiative of the Canadian Water Network – Oil Sands, is principal investigator for an Alberta Environment funded study of process-affected water and groundwater in the oils sands region, leads three industry funded projects on SAGD water development, and has participated in eight regional critical acid loads studies across Canada including work carried out by the NOxSOx Management Working Group of CEMA and the Canadian Council of Ministers of the Environment Acid Rain Task Group.

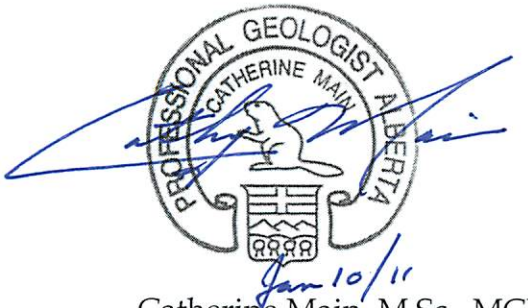
Dr. Kelly Munkittrick Ph.D., is the Associate Director of the Canadian Rivers Institute and holds a Tier 1 Canada Research Chair in Ecosystem Health Assessment at the University of New Brunswick in Saint John, New Brunswick, Canada. He is also the Program Leader for the Watersheds and Ecosystems Theme of the Canadian Water Network, is Co-leader of the Lakes Working Group for the GEF IW:Science project executed by UNU-INWEH, and on the Great lakes Fisheries Commission Board of Technical Experts where he leads the theme on Ecosystem Dysfunction. Prior to his appointment at UNB, he worked for 11 years for the Canadian 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 Lakes Laboratory for Fisheries and Aquatic Sciences. His research interests are related to assessing the environmental impacts of industrial and agricultural activities, and on developing methods for environmental effects monitoring and for the cumulative effects assessment of multiple stressors on aquatic environments. He currently have active projects on assessing environmental impacts in Sri Lanka, Bhutan, Chile, Uruguay, Brazil, Cuba, the US and Canada, and has worked, taught or given invited lectures in more than 25 countries.

Dr. John Post Ph.D., is a Professor of Ecology and Evolutionary Biology at University of Calgary, Calgary, Alberta, Canada. He received a Ph.D. from York University in Toronto in 1987 in fish ecology followed by a post-doctoral fellowship at University of Wisconsin and a term faculty position at University of British Columbia before being appointed in the Department of Biological Sciences at the University of Calgary in 1991. John's research spans fundamental fish population ecology, climate change biology, conservation biology, fish habitat requirements and harvest dynamics of freshwater fishes. A current research focus involves developing models and field tests that integrate hydraulics and fish habitat requirements at local and watershed scales to

provide predictive tools for determining instream flow needs for viable river ecosystems. John and his students use observations, experiments and models to understand ecological processes in structured fish populations, harvest dynamics and habitat requirements to maintain viable fisheries. John and his students have published over 100 papers, book chapters and technical reports in fish ecology and fisheries.

Dr. Shaun Watmough, Ph.D., (BSc. Applied Biology, Liverpool Polytechnic, U.K., Ph.D. Plant Stress Physiology, Liverpool John Moores University, U.K.) is an Associate Professor in the Environmental Resource Science Program at Trent University in Peterborough, Ontario. His research focuses on ecosystem environmental stress, and his research interests include forest ecology, biogeochemistry, forestry, air pollution, climate change, trace metals, eutrophication, and environmental modeling. Dr. Watmough and his students have published more than 80 papers in international peer-reviewed journals and he has contributed to two chapters in the 2004 Acid Rain Assessment produced by Environment Canada. Dr Watmough has been involved in acidification research and critical load assessments in Ontario, Nova Scotia, British Columbia, Manitoba and Saskatchewan for more than a decade. Dr. Watmough has also worked in the Athabasca Oil Sands on several projects funded by Wood Buffalo Environmental Association (WBEA) and Cumulative Environmental Management Association (CEMA) to study and model the impacts of acid deposition on the terrestrial and aquatic environments. This research has resulted in a number of peer-reviewed publications.

Respectfully submitted,



Catherine Main, M.Sc., MCIP, RPP., P.Geol., P.Geo.
Program Leader
Integrated Water Management Program

REFERENCES

Ayles, Burton Monique Dubé and David Rosenburg. 2004. Oil Sands Regional Aquatic Monitoring Program (RAMP) Scientific Peer Review of the Five Years Report (1997 – 2001) Prepared for and submitted to the RAMP Steering Committee.

RAMP, 2009. RAMP Technical Design and Rationale Document. Prepared for RAMP Steering Committee by Hatfield Consultants, Kilgour & Associates Ltd., Klohn Crippen Berger Ltd. and Western Resource Solutions. December 2009.

RAMP, 2009. Regional Aquatics Program – 2009 Technical Report. Hatfield Consultants. West Vancouver, British Columbia..

Regional Aquatic Monitoring Program (RAMP). 2005. RAMP: Technical design and rationale – Final Report. Hatfield Consultants. West Vancouver, British Columbia.

Smith, Eric P. 2002. BACI Design. Encyclopedia of Environmetrics: Volume 1. Edited by Abdel H El-Shaarawa. and Walter W. Piegorsch. John Wiley & Sons, Ltd.: Chichester, pp.141 – 148.

Whittier and Hughes. 2008. An Assessment of the Regional Aquatic Monitoring Program (RAMP) Fish Survey. Prepared for Hatfield Consultants Partnership. Department of Fisheries and Wildlife. Oregon State University, Corvallis, Oregon.

APPENDIX A

Dr. Donald Burn Review

PEER REVIEW OF THE CLIMATE AND HYDROLOGY COMPONENT OF THE REGIONAL AQUATICS MONITORING PROGRAM (RAMP)

DONALD H. BURN

DEPARTMENT OF CIVIL AND ENVIRONMENTAL ENGINEERING

UNIVERSITY OF WATERLOO

INTRODUCTION

This report summarizes the peer review for the Climate and Hydrology component of the Regional Aquatics Monitoring Program (RAMP). While the primary focus is on the Climate and Hydrology component of the program, the review also includes comments on other components of the RAMP, particularly as they relate to the Climate and Hydrology component.

REVIEW APPROACH

This report is primarily based on a review of the following documents and information sources:

1. The 2009 Design and Rationale Document;
2. The 2009 Technical Report (reports from select earlier years were perused as well, but in much less detail);
3. The RAMP Terms of Reference;
4. The RAMP members' website;
5. The RAMP database; and
6. The Scientific Peer Review of the Five Year Report (1997-2001).

In all cases, the main focus of the review, and hence the comments that follow in later sections of this report, was on the sections that relate to the Climate and Hydrology component of RAMP with lesser attention paid to the other components. The rationale for this approach was to utilize the expertise of the reviewer on areas of greatest knowledge and experience. As well as focusing on the Climate and Hydrology component, the review additionally looked at areas where the Climate and Hydrology component does, or should, interact with other components of RAMP.

COMPONENT REVIEW

STRENGTHS OF EXISTING PROGRAM

Monitoring in a northern environment in Canada is generally a challenging undertaking. In addition to the obvious access issues due to the remote location of monitoring sites and the difficulties of severe weather conditions, analysis of climatic and hydrological variables is further challenged by the short data records that are available for most of the gauging stations in Canada's north. Lengthy data records are essential for identifying, using statistical tests, changes that may have occurred, or may be occurring, in hydroclimatic variables of interest. The integration of existing hydroclimatic monitoring stations with the new RAMP monitoring stations has been essential to ensure that the database of hydroclimatic variables includes record lengths that are sufficiently lengthy to identify changes that may be occurring in the variables of interest.

Joint monitoring at some locations by RAMP and the Water Survey of Canada (WSC), where the RAMP monitoring is being used to collect data during the winter season when WSC does not collect data at all locations, is an additional strength of the monitoring program. The joint monitoring has the potential to provide valuable information regarding winter time (generally low flow) conditions that would not otherwise be available for some locations. However, as discussed further below, there are some locations where the monitoring is still only conducted on a seasonal basis and there is a loss of information associated with this monitoring strategy.

POTENTIAL AREAS OF IMPROVEMENT TO PROGRAM

Potential areas for improvement are outline in this section and expanded upon in the Discussion section that follows.

WATER BALANCE MODELLING

The water balance model is the main basis for comparing baseline and test conditions for a watershed that has undergone, or is undergoing, oil sands development. The water balance model estimates what the baseline conditions for a watershed would have been using daily

streamflow data collected at a test gauging station with adjustments applied to the test condition data to reflect:

1. Any industrial withdrawals and discharges;
2. Areas of the watershed that have been clear cut, from which greater runoff production is expected; and
3. Areas of the watershed that no longer contribute flow to the watershed outlet (closed-circuit areas).

The water balance model that is used is a very simple modelling approach that has several important assumptions, either implicit or explicit, which likely impact the quality of the results obtained from the model. Some of the more important model assumptions, and their potential implications, are outlined below.

Daily withdrawal and discharge data are used in the model, which operates on a daily time step, if daily withdrawal/discharge data are available. However, for some withdrawals and discharges, the data are only available on a monthly or, in some cases, on an annual basis. If withdrawal/discharge data are not available on a daily basis, the assumption is made that the rate of withdrawal or discharge is constant throughout the time interval (month or year) for which withdrawal/discharge data are available. The impacts of this assumption on the modelling results will depend on the actual amount of variability in the withdrawal and discharge rates and will be most critical in the low flow season when the withdrawals and discharges represent a larger percentage of the total flow. It is not possible to precisely quantify the impacts of this assumption, but the impacts could be substantive, particularly in the low flow season, and could therefore result in estimates of baseline measurement endpoints that are in error.

The modelling approach implicitly assumes that the rate of runoff production is constant from all undisturbed portions of the watershed (i.e., a lumped modelling approach). In reality, there will be spatial variability in runoff production reflecting different combinations of ground surface slope, soil cover and type, and land cover as well as spatial variations in precipitation. The impacts of this assumption will be greatest in watersheds that are heterogeneous and will likely be more pronounced on larger watersheds than on smaller watersheds.

The approach for representing the changes in runoff that result from closed-circuit areas and from clear cut areas assumes that the location of the closed-circuit or clear cut area does not have any impact on the runoff that arrives at the watershed outlet. The modelling thus does not have the capability to capture the more complex flow dynamics that will undoubtedly occur as a result of changes in the flow pathways resulting from closed-circuit

or clear cut areas. The magnitude and timing of runoff changes in an altered watershed will actually depend on the location of the altered area within the watershed (i.e., headwater locations versus locations closer to the watershed outlet). This behaviour is not reflected in the water balance model, which in essence assumes that the altered areas are distributed evenly throughout the watershed.

The increase in runoff from clear cut areas is assumed to be a constant 20%. In reality, the change in runoff production resulting from clear cutting is likely to be a function of the time of the year (changing on a seasonal basis) and will vary in magnitude from one location to another. Since clear cut areas are largely a transition land classification, the impacts of this assumption are not likely overly problematic, but could affect the estimation of the baseline values of the low flow related measurement endpoints.

The water balance model clearly plays a pivotal role in the evaluation and quantification of the effects of oil sands development activities on the hydrology of watersheds in the study area. Therefore, if the water balance model gives poor results, the evaluation of the impacts of oil sands development will also be in error. If the water balance model is to be the basis for the evaluation of impacts, it is important that the modelling approach be as accurate as possible. This is particularly true for the low flow and winter season discharge measurement endpoints due to the greater sensitivity of these measurement endpoints to the modelling assumptions of the water balance model.

Groundwater Monitoring

There is, at present, no groundwater monitoring as a formal part of RAMP, although it appears that there are some groundwater monitoring activities that are conducted in the study area. The lack of groundwater monitoring as a formal part of RAMP is a serious limitation of the monitoring program and results in a lack of information on some of the impacts of oil sands development activities within the study area. There are development related groundwater withdrawals occurring in the study area that can be expected to have an impact on the surface and subsurface water levels in the area. It will be very difficult to evaluate these impacts on a regional basis without including groundwater monitoring in RAMP.

Groundwater is an important determinant of the low flow conditions in a water body such that changes in groundwater can have an effect on the magnitude and timing of low flow conditions. This is important for the Climate and Hydrology component, where low flow is a measurement endpoint, but also for aquatic organisms, which can be severely impacted by alterations in low flow conditions. It is hard to understand how these important aspects of the hydrological regime can be sufficiently understood without monitoring groundwater conditions in the study area.

An understanding of the nature of the changes in some water quality variables requires an understanding of the sources and pathways for waters in various surface water bodies. The acidification of lakes is also influenced by the source of water to the lake (i.e., surface water versus groundwater source). Characterizing these sources and pathways requires monitoring of groundwater to understand, and possibly attribute, changes that may occur in water quality variables.

NETWORK DESIGN

The monitoring network for Climate and Hydrology has evolved over time and continues to evolve, especially as baseline gauging stations change to test gauging stations as a result of the development of oil sands projects upstream of the existing (baseline) gauging stations. While the monitoring network is changing, the changes that are occurring are largely reactive as opposed to proactive. A more systematic network design process is needed to obtain the database of long term monitoring records required to identify changes and shifts in the hydrological regime. The network design process needs to anticipate the development of oil sands properties and locate gauging stations both upstream and downstream of potential development sites to ensure that:

1. Baseline conditions are continually monitored for the undisturbed portion of the watershed; and

2. Downstream baseline record lengths for the period prior to development are sufficiently lengthy to form a strong basis of comparison with test conditions measured after the development of oil sands in the watershed.

Consideration should be given to the concept of developing a "regional" network of gauging stations, consisting of both stations in the study area and stations close to the study area. The latter should be stations that can be considered to be hydrologically similar to the stations within the study area. This network of stations could provide a baseline basis of comparison with which to evaluate the measurement endpoints from the test stations. This approach will allow test stations to be compared to one or more "similar" baseline stations and the response of the test station in a particular year to be evaluated against the corresponding baseline station(s). An approach such as this could minimize the need for the water balance approach that is currently the key component for comparing baseline and test conditions.

The streamflow gauging stations need to be monitored on a year round basis, rather than on a seasonal basis, which is currently the practice for some (especially smaller) watersheds. There are two main reasons for monitoring on a year round basis. First, one of the measurements endpoints is the mean winter discharge, which obviously cannot be evaluated without year round monitoring activities. Second, there are several watersheds where in one or more years the monitoring appears to have started too late in the season to

capture the entire spring freshet, which is an important part of the hydrological regime from a water quantity perspective and also from an ecosystem perspective. There are several example watersheds where either part or the entire spring freshet appears to have been missed (in at least some years), including: the Tar River; the Calumet River; the Ells River; Poplar Creek; and Fort Creek.

MEASUREMENT ENDPOINTS USED

At present, there are four measurement endpoints for the Climate and Hydrology component:

1. The mean open water season discharge;
2. The mean winter discharge;
3. The annual maximum daily discharge; and
4. The open-water season minimum daily discharge.

The present list of measurement endpoints is clearly limited and does not encompass the full array of measurement endpoints that have been used in the oil sands project EIAs. It will thus be difficult to determine the extent to which some of the projected impacts are in fact occurring. Measures related to the timing of low flows, the timing of high flows, the onset of the spring freshet, etc. could be readily added to the list of measurement endpoints using the streamflow data presently available and the existing monitoring network. Measurement endpoints derived from timing measures would provide a more complete assessment of the changes that may be occurring in the hydrological regime of the study area.

The 2009 Design and Rationale Document indicates that additional measurement endpoints related to hydrological extremes will be able to be added in future years when additional years of data are available. Expanding the list of measurement endpoints in the manner described in the 2009 Design and Rationale Document will be challenging mainly due to the fact that existing baseline stations in many locations are becoming test locations which means that the available data record for most baseline and test locations will remain short for the foreseeable future. It may be necessary to rethink the overall strategy of determining differences between baseline and test conditions in order to obtain a more comprehensive set of measurement endpoints.

TREND ANALYSIS

There are several components of RAMP that involve the calculation of a trend over time in a measurement endpoint. An example of this is the Acid Sensitive Lakes component where the Mann-Kendall trend test is used to evaluate changes over time in numerous measurement endpoints related to the acidification of lakes. According to the 2009 Design

and Rationale Document, the record length criteria used to implement the Mann-Kendall test is a record length of at least seven years. The power of the Mann-Kendall test increases with the record length and will generally be very small for a record length of only seven years. This implies that the results of the trend test on a very short record can be quite misleading. The fact that a significant trend is not identified does not imply that a significant trend does not exist as there may be a trend but the trend test is not sufficiently powerful to identify it at the selected significance level. There are also likely to be many cases where a significant trend is identified at one point in time, but the collection of additional years of data may mean that the trend is no longer significant in subsequent years. There are several examples of this latter type of behaviour in the 2009 Technical Report. This can also result in misleading conclusions. A better approach to address changes in variables when the record length is so short is to summarize the slope values (calculated using a robust estimate of the slope) and highlight changes that are, and are not, consistent with the hypothesis of interest (increased acidification of lakes, in this case). This approach would remove the need to establish statistical significance and instead focus on the weight of evidence in support of, or contrary to, the basic hypothesis that is being evaluated.

DISCUSSION

RELEVANCE OF MONITORING APPROACH

The monitoring activities within the Climate and Hydrology component of RAMP involve the collection of:

1. Climate data at climate stations;
2. Snow course survey data;
3. Lake levels;
4. Streamflow data at hydrometric gauging stations; and
5. Climate data collected at hydrometric stations.

The monitoring activities are thus comprehensive and somewhat unique for a remote and northern environment. However, the question remains as to whether the current analysis of the data that are collected can identify changes in the hydrological regime within the study area.

A major limitation of the analysis approach used in the Climate and Hydrology component is the water balance model, which, as noted above, has numerous assumptions that may prove to be problematic. As outlined above, some of these assumptions could result in the

estimates of the baseline measurement endpoints for disturbed watersheds being in error. This calls into question the capability of the monitoring program to identify all of the relevant effects of the oil sands projects. The nature of the modelling approach is such that the confidence in estimated changes in some measurement endpoints will be greater than the confidence in the estimated changes in other measurement endpoints. For example, the estimated changes in the mean open water season discharge are likely to be fairly reliable since the impacts of the various assumptions in the water balance model will be mitigated by the averaging of daily discharge values over a multi-month period. However, the estimated changes in the remaining measurement endpoints could be in considerable error.

It is not clear that the monitoring and analysis strategy currently in place will lead to an accurate determination of any long term trends or changes that may be occurring. This is due in part to the limitations of the network design, which results in very few long term data records from which statistically significant trends or changes can be identified (i.e., most of the baseline and test locations have very short data record lengths). A second limitation is the calculation of baseline measurement endpoints using the water balance model, as discussed above. The modelling and analysis approach adopted will make it very difficult to distinguish between trends and changes that arise from the oil sands activities and trends and changes that are occurring from, for example, climate change. It is possible that the oil sands activities in the study area will either exacerbate or partially mitigate the impacts of climate change, although the former would appear to be the more likely outcome.

GAPS IN MONITORING APPROACH

The main gap in the monitoring approach, as noted above, is the lack of groundwater monitoring as a part of RAMP. This deficiency could be alleviated by implementing a groundwater monitoring program and/or utilizing groundwater data that are currently collected in the study area by others.

The calculation of additional measurement endpoints would be beneficial to provide a more complete understanding of the hydrological regime, and changes to the regime, in the study area. There may be some challenges in effectively estimating measurement endpoints related to the timing of runoff responses, as discussed above, using the water balance model (see the discussion of the limitations of this model presented above).

LINKAGES AND INTEGRATION WITH OTHER PROGRAM COMPONENTS

The data collected from the Climate and Hydrology component is of use in many of the other RAMP components to help place the monitoring results in context. However, the results from the other RAMP components do not assist with the assessment of the measurement endpoints for the Climate and Hydrology component. There does seem to be

a fair bit of harmonization of data collection locations; whether the current amount of data collection harmonization is sufficient can best be determined in the context of the requirements of the other RAMP components.

At present, RAMP makes fairly limited use of the climate data that are collected. Climate data are used in the annual Technical Reports to characterize for a given year the overall hydroclimatic conditions for the study area, and to a lesser extent for individual watersheds. This characterization is in terms of a dry year, an average year or a wet year, etc. This characterization is useful to help understand and interpret the measurement endpoints for the subject watersheds. However, more could be done with these data, particularly if the Climate and Hydrology component were to move beyond the simple water balance model to the use of a hydrological model to better represent the response of watersheds and the effects of oils sands development on the watersheds.

RECOMMENDATIONS

The implementation of the following recommendations, also summarized in the Appendix, could improve the evaluation of the impacts of oils sands development:

1. Reconsider the use of the current water balance model as the basis for estimating baseline measurement endpoints for comparison with observed test measurement endpoints.
2. Incorporate groundwater modelling as a formal part of the Climate and Hydrology component of RAMP.
3. Develop a more proactive approach to data collection network design.
4. Where practical, monitor streamflow gauging stations on a year round basis.
5. Add additional measurement endpoints that reflect the timing of the hydrologic response of a watershed.
6. Reconsider the use of trend analysis for record lengths that are very short.

APPENDIX

Table 1 Summary of recommendations

Component	Issue	Recommended Change	Rationale
Climate and Hydrology	Water balance model	Reconsider the use of the current water balance model as the basis for estimating baseline measurement endpoints	There are numerous assumptions inherent in the water balance model that likely have a negative impact on the estimates of baseline conditions
Climate and Hydrology	Groundwater monitoring	Incorporate groundwater modelling as a formal part of RAMP	Many measurement endpoints are affected by groundwater conditions yet these are not measured as a part of RAMP
Climate and Hydrology	Network design	Develop a more proactive approach to data collection network design	The current network design process is largely reactive and is thus less efficient than it could be
Climate and Hydrology	Streamflow measurement	Monitor streamflow gauging stations on a year round basis	There are some watersheds where an important component of the runoff regime is missed due to seasonal monitoring of streamflow
Climate and Hydrology	Measurement endpoints	Add additional measurement endpoints that reflect the timing of the hydrologic response	Timing measures are important for determining the impacts of development
Acid Sensitive Lakes	Trend analysis	Reconsider the use of trend analysis for record lengths that are very short	Trend analysis for a very short record length can be misleading

APPENDIX B

Dr. John Gibson Review

Reviewer: John Gibson, Alberta Innovates

Date Submitted: 25 September 2010

Introduction

A review of RAMP's Climate and Hydrology component was requested by Catherine Main, Alberta Innovates in June 2010 on behalf of the RAMP Steering Committee. Dr. Gibson attended a coordination meeting held at Hatfield Consultants in North Vancouver on 15 July 2010, attended by members of the review panel. Dr. Gibson accepted the task of reviewing the Climate and Hydrology component as well as the additional task to review the Acid Sensitive Lakes component. General comments on other components are also included. These reviews are presented in the following document.

Review Approach

Accordingly, this review aims to provide an assessment of two main components of the RAMP program, namely : (i) Climate and Hydrology, and (ii) Acid Sensitive Lakes. The review is intended as a scientific appraisal of major issues such as design and rationale, groundwater monitoring, methodology, and data presentation and accessibility. As per RAMP's Technical Design and Rationale, Section 1.1.1 Item 9, this document is also structured as a review of the program's objectives against its results, and includes recommended adjustments necessary for the program's success. The scope of the review is defined largely by the documents and resources reviewed including:

- RAMP Technical Design and Rationale, December 2009, particular emphasis on the following material:
  - Sections 2-0 to 2-23 "Results of EIA review"
  - Sections 3-1 to 3-6 "Ramp Design and Rationale",
  - Sections 3-4 to 3.5.7 "Climate and Hydrology", and
  - Sections 3.9 to 3.9.6 "Acid Sensitive Lakes"
- RAMP 2009 Technical Report Draft, April 2010, with particular emphasis on the following material:
  - Sections 3-1 to 3-6 " 2009 Ramp Monitoring Activities"
  - Section 4.0 " Climatic and Hydrologic Characterization of the Athabasca Oil Sands Region in 2009"
  - Section 5.0 " Results for Individual Watersheds"
  - Section 6.0 " Special Studies"
  - Section 7.0 "Regional Synthesis"
  - Section 8.0 "Conclusions and Recommendations"
  - RAMP 2008 Final Report, April 2010, with particular emphasis on the following material: Climate and Hydrology and Regional Studies
- 2005-2008 Technical Reports, with particular emphasis on Climate and Hydrology and Acid Sensitive Lakes
- Other reports: Timoney Final Report Nov07.pdf, AENV Water for Life Technical Report Oct07.pdf, NREI Synthesis report.pdf, RAMP Fish Survey Assessment_Bob Hughes.pdf, TarSands_Environmental Defence Report_Feb2008.pdf ; Timoney Open Conserv Biol J paper.pdf, Kelly -

Schindler et al 07Dec09 PNAS papers, Review of Kelly et al paper (V3_01Feb10).pdf
- Other relevant reports and references including WRS 2004, WRS 2006
- Online database and query

For the acid sensitive lakes component, this review also draws on first-hand knowledge gained from participation in CEMAs NSMWG sponsored research projects and similar programs sponsored by Environment Canada and the Canadian Council of Ministers of the Environment Acid Rain Program. This scientific perspective is provided in a series of recent publications including:

- Scott, K.A., Wissel, B., Gibson, J.J., Birks, S.J., 2010. Limnological characteristics and acid sensitivity of boreal headwater lakes in northwest Saskatchewan, Canada. Journal of Limnology 69 (Suppl. 1) 33-44, 2010 - DOI: 10.3274/JL10-69-S1-05.
- Jeffries, D.S., Semkin, R.G., Gibson, J.J., Wong, I., 2010. Recently surveyed lakes in northern Manitoba and Saskatchewan, Canada: characteristics and critical loads of acidity. Journal of Limnology 69 (Suppl. 1) 45-55, 2010 - DOI: 10.3274/JL10-69-S1-06.
- Gibson, J.J., Birks, S.J., McEachern, P., Hazewinkel, R., Kumar, S., 2010. Interannual variations in water yield to lakes in northeastern Alberta: Implications for estimating critical loads of acidity. Journal of Limnology 69 (Suppl. 1) 126-134, 2010 - DOI: 10.3274/JL10-69-S1-13.
- Gibson, J.J., Birks, S.J., Jeffries, D.S., Kumar, S., Scott, K.A., Aherne, J., Shaw, P., 2010. Site-specific estimates of water yield applied in regional acid sensitivity surveys across western Canada. Journal of Limnology 69 (Suppl. 1) 67-76, 2010 - DOI: 10.3274/JL10-69-S1-08.
- Bennett, K.E., Gibson, J.J., McEachern, P., Water yield estimates for critical loadings assessment: comparisons of gauging methods vs. an isotopic approach, Canadian Journal of Fisheries and Aquatic Sciences 65: 83-99..
- Gibson, J.J., Prepas, E.E., McEachern, P., 2002. Quantitative comparison of lake throughflow, residency, and catchment runoff using stable isotopes: modelling and results from a survey of boreal lakes. Journal of Hydrology 262: 128-144

Assessment of Program's Results Against Objectives

Given that the overall task of this review is to assess the results of the program against the objectives it was considered important to provide discussion in this context.

The Overall RAMP objectives, taken from RAMP Technical Design and Rationale, December 2009, Section 1.1.1 are provided followed by review comments on each issue, as follows:

Comments	Component	Recommendation
1.1 Monitor aquatic environments in the oil sands region to detect and assess cumulative effects and regional trends		
The streamflow measurement program is a valuable monitoring activity that approaches conformity with the Water Survey of Canada operational standards and effectively interfaces with the provincial/national operated monitoring programs in the area. Stations discontinued by Water Survey of Canada have been re-established by RAMP which has helped considerably to maintain continuity of gauging in the region. RAMP has also complemented the WSC efforts by adding winter gauging of low flows, an important control on water quality, benthic invertebrates and fish habitat. RAMP stations still include year-round and seasonal gauging stations. Seasonal gauging is the only practical method for gauging in many small tributaries due to icing of culverts and channel overflow. Some gauging stations have significant in-channel vegetation and less than ideal gauging conditions due to beaver dams, aquatic vegetation and poorly defined stream channels, as noted in Appendix 4 Design and Rationale. Relative quality of gauging stations/records needs to be reported using the WSC flags.. When spot measurements are made in winter it is very important to include "B" flags as appropriate to indicate ice-affected conditions. Overall, operation and coordination of monitoring at 31 gauging stations (18 year-round and 13 seasonal) is a substantial contribution to understanding of runoff variability in the region. Continual reduction in baseline stations over the past decade has occurred with about 7 baseline stations changing from baseline to test. Three more are expected to become test stations in the near future as new developments proceed. The gauging network augmented by Water Survey of Canada's network, is considered to be a suitable basis for assessment of natural and disturbance-related streamflow characteristics in rivers. It is important to note that detection of changes in streamflow even at locations with long gauging records can be complicated in snowmelt and wetland dominated sites due to interannual variability in runoff response. Often 10 to 30 years of record are recommended to adequately characterize normal range of natural variability in streamflow. Due to continual reduction in baseline stations, special attention needs to be given to maintaining as many stations as possible over the next four decades to monitor	CH CH	#1

anticipated cumulative changes in regional runoff response.		
The stream gauging network is considered to be adequate for detection of water quantity/quality changes in downstream areas, but upstream hydrologic conditions in developing areas including lakes, wetlands, and groundwater is not adequately captured by the current monitoring to associate a change with a specific impact cause, pathway, development, or operator.	CH	
Water level variability monitoring in lakes is much more limited that streamflow monitoring, and includes only two baseline sites (McClelland and Kearl Lakes) and one test site (Isadore's Lake). Note that Shipyard Lake is not listed obviously in Table 3.4.Design and Rationale nor is it available for download online (unless by another name). The lake level network is considered inadequate for capturing spatial variability in storage changes, and includes too few stations to adequately define regional trends in headwater lakes and higher order systems. The network would also benefit from additional water level stations in lakes, wetlands and shallow soils, to provide information on seasonal and interannual storage changes across the region.	CH	
The RAMP climate monitoring network consists of only a few stations. The Hatfield Team, apparently recognizes the limitations of the climate network south of Fort McMurray, and proposes addition of a new climate station in this area. Climate data are limited to basic parameters such as temperature and precipitation but offer insufficient spatial coverage to monitor either regional shifts at the large scale or watershed-scale differences at the small scale. Important climate-driven processes such as evaporation and transpiration cannot be easily calculated from RAMP monitoring data despite their importance for hydrological modeling. Snowpack monitoring as summarized in Fig. 4. 1-4 is considered to be generally sufficient for the purpose of estimating water inputs to the hydrological system but complimentary sampling of PAHs and heavy metals would be a prudent addition to the network given the results from recent studies (eg. Kelly et al. 2009, 10). Samples are already informally collected to characterize the stable isotope content of the snowpack (^{2}H,^{18}O) which is useful for calibration of hydrological models.	CH, WQ	#2
Addition of complimentary water table monitoring stations in the major terrain units, i.e. low-lying areas, mixed deciduous, jack pine and open land (wetland)/lake would be of great value for hydrological modeling, but this may be better addressed by sideline research projects rather than by expansion of core monitoring.	CH	

Overall, the program is not currently structured in a stand-alone capacity to detect and assess cumulative effects and regional trends, although some modifications to the monitoring design could significantly improve the effectiveness of the program and network so that they approach these objectives, at least from the climate and hydrology perspective. In particular, incorporation of groundwater monitoring is required to truly assess cumulative effects on the hydrological cycle.	CH, All	
Oil sands extraction is a groundwater intensive industry with development of groundwater supplies to significantly exceed surface water supplies by 2 to 3 times as in situ deposits are developed. While the current water balance approach for testing impacts, using closed-circuit areas with zero runoff and non-closed circuit areas with 20% runoff, may be sufficient at present to simulate and compare impacts of mining- and development-related disturbances on runoff in smaller watersheds, the groundwater system will likely be affected more extensively on an interbasinal, regional scale in future. Surface/groundwater exchange has a profound impact on discharge, water quality, and other factors and these connections may be locally strong or weak. Weak connections lead to decoupling of the surface and groundwater systems while strong connections ensure good connection, and a coupled response.	CH, WQ	
Groundwater seeps are a prominent feature of the incised river courses in the area. Mapping of groundwater seeps using EM resistivity surveys is one potential method for defining the location of natural or anthropogenic groundwater inflows, for inventorying PAH, naphthenic acids, priority pollutants and for gaining a better understanding of the influences of various groundwater units/formations on water quality along specific river reaches. Inflow from such seeps is believed to be a larger influence on reach water quality than so-called "contact" between the river water and geological formations, as these materials have often been effectively leached over hundreds or thousands of years.	WQ, CH	
Other related industry in the area has also impacted the groundwater system which is of importance for evaluating cumulative impacts. For example natural gas development has already had a profound impact on groundwater levels in the southern Athabasca region, with water level drops in the 10s of metres over the past 40 years in some areas (Gordon Lake area), and this impact is part of the cumulative impact in the region. Injection of saline wastewater and potential leakage to surface waters will be another potential impact on aquatic ecosystems in future.	CH	

Groundwater water test data and water quality data are routinely collected by operators for compliance with AENV regulations. Due to importance of groundwater data for use in assessing possible causes of changes in WQ data it is recommended that some of this information be made available via RAMP. A reasonable approach would be to make water level and water quality data available for selected wells in each formation (total of ~20-30 wells).Groundwater monitoring may therefore be largely accomplished by data sharing with Operators, and via the Alberta Environment GOWN monitoring network.	CH, WQ	#3
EM terrain conductivity mapping of river reaches and sampling of prominent seeps to establish baseline natural inflows and differentiate from possible process-affected water inflows either present or occurring in future. Such surveys might be coupled to reach-wise synoptic sampling surveys for isotopes and geochemistry to identify points or zones of inflow/potential outflow.	WQ, CH, All	#4
The methodology used to reconcile water balance of test/baseline reaches of the network, i.e. $Hyd = Hyd + I - I + R - R$ where: *HydB* is the *baseline* hydrograph for 2009; *HydO* is the *test* hydrograph which was observed in 2009; *Iw* are the focal project withdrawals from the watershed; *Ir* are the focal project releases to the watershed; *Rn* is the natural runoff that would have occurred in the watershed, but was intercepted or closed-circuited by focal projects in 2009; and *Ri* is the incremental increase in runoff caused by land cleared within the Basin, is utilized as a primary basis for establishing and quantifying hydrological impacts.	CH	
While this approach serves as a useful first-approximation of naturalized flow, and for identifying unexpected development impacts, there are many assumptions used in the comparisons that are admittedly based on the professional judgment of the Climate and Hydrology Component subgroup under the RAMP Technical Program Committee, and.may not be valid for some times/locations. Such assumptions include use of constant runoff for undisturbed areas, use of 20% higher runoff for cleared areas, and use of zero runoff for closed-circuited areas. These values do not account for expected changes in runoff due to seasonal or interannual variations in antecedent soil moisture conditions, or other modifying factors, and may complicate establishment of some measurement endpoints,	CH	

particularly low flows. This approach excludes influences from groundwater inputs to surface water or groundwater capture by SAGD development. It also does not address changes in catchment responsiveness caused by changes in catchment area or disruption of flow pathways across disturbed sites. Predictions of hydrological changes using these simple formulae will expectedly become more difficult as the scope of development increases due to cumulative effects.	All	#5
Standard water quality analysis should include Petroleomics (FT-IRCMS) to scan for natural organic compounds as a complimentary technique to naphthenic acid partitioning. Petroleomics work to date has suggested that up to 5000+ distinct organic compounds may be present in some natural groundwaters in the oil sands region, and these may be used as fingerprint of water origin. Background on this technique can be found in Rodgers et al. (2005). Petroleomics: Ms returns to its roots, *Anal. Chem.* **2005**, 77, 20A.	CH, WQ, ASL	#6
There is an obvious lack of connection between the climate, hydrology and water quality network and the acid sensitive lakes network. Addition of any future baseline streamflow stations or lake level stations might consider seeking suitable locations in the acid-sensitve lakes watersheds. Addition of one interannual lake level station in each of the six acid sensitive lakes groups would be a significant contribution to both the Climate and Hydrology and Acid Sensitive Lakes component objectives.	CH, ASL	#7a
1.2 Collect baseline data to characterize variability in the oil sands region Due to pace of expansion in regional development, careful attention to maintaining the Baseline/Test balance of stations, including selection of additional baseline stations perhaps even outside the Regional Municipality of Wood Buffalo will be required. Test stations are not expected to show the same variability as baseline stations for obvious reasons. Increasing reliance on naturalized flow records for basins as baseline stations shift to test stations is also discouraged as these stations will not likely provide sufficiently accurate information for peak flow or winter low flow assessments. Statistical trends in streamflow need to be assessed both for baseline and test stations.	All	

Selection of new baseline hydrology sites should consider situating the stations to support the acid-sensitive lakes component, even if outside the vicinity of development or even the regional municipality of wood buffalo. They should ideally not be slated for development for 20+ years.	CH,WQ, ASL	#7b
Addition of special projects such as the Nexen lakes to the overall RAMP monitoring framework would likewise be a straightforward approach to improving coverage of sites south of Fort McMurray.	CH, WQ, ASL	#8
1.3 Collect and compare data against which predictions contained in the EIAs can be assessed The 2009 RAMP Design and Rationale document, Appendix A3 presents a summary of impact pathways extracted from EIAs. The summary identifies 809 impact scenarios from 17 EIAs and assigns endmenbers from which to assess impacts. From this analysis 378 impact pathways directly relate to hydrology, and 247 indirectly relate to hydrology. This implies 625 of 809 (77%) of impacts are via hydrologic pathways, which reinforces the importance of hydrologic monitoring conducted by RAMP. It is reasonable to define qualitative criteria for assessment based on percentage change in flow/level from the measurement endpoints. For hydrology, areas with open pit mining are apparently treated differently than SAGD. Mining impacts considered as follows: 5% (negligible), 10% low, 10 to 30% moderate, and >30% high. For SAGD, negligible to low is <1%, moderate is 1 to 10% and high is >10%, which appears to recognize more pronounced impact of mining on the near-surface water cycle. Section 3.1.7.5 of the 2009 Technical Report states: "The percent difference between the *test* and *baseline* values of the hydrologic measurement endpoints were used to classify results as follows: ± 5% - Negligible-Low; ±5-15% - Moderate; > 15% - High. These ranges were derived from criteria for determining effects on hydrologic measurement endpoints in a number of EIAs prepared for oil sands projects (RAMP 2009b)." This appears to differ from the differential percentage scales presented in Table 2.12 of the Technical Design and Rationale as some projects, namely CPC Surmont, Syncrude Aurora and Husky Thermal Sunrise set lower limits. The rationale for this needs to be addressed particularly since it seems that the insitu projects have used lower thresholds for defining impacts. The network is structured appropriately for comparison with watershed-scale riverine endpoints, as outlined in most of the EIA predictions, provided that river monitoring is continued and that the	CH, All All	#9

baseline/test balance is preserved. The only predictions that may be difficult to answer conclusively will the predictions related to groundwater impacts on flow, water quality, fish and benthic invertebrates. These include 78 of 809 EIA scenarios related to hydrological impacts, including: nos. 109,110,111,112,113,114,116,163,164,195,196,197,198, 199,200, 201,202, 203, 208, 209; water quality impacts, including 653,655,660, 661, 668, 669,671, 686,687,689,690,692,693, 695, 696, 698, 699, 700, 701, fish/fish habitat and benthic invertebrate impacts, including: 275, 276, 277, 278, 295, 296, 297, 302, 303, 304, 310, 311, 312, 320, 344, 345, 347, 348, 350, 351, 353, 354, 356, 359, 360, 362, 363, 365, 368, 369, 468, 512, 526, 527, 528, 531, 532, 533 and 540. There are potential problems with use of a surface water focused assessment endpoints when looking at groundwater impacts. In short, surface/groundwater connections and/or exchange may be locally strong or weak. Weak connections can lead to decoupling of the surface and groundwater systems while strong connections ensure good connection, and a coupled response. In the case of areas that are decoupled, the impact of groundwater abstraction may be advanced before the assessment endpoint is affected. In general, for the currently used scenarios the case of zero impact can be tested, but it may be difficult to attribute an observed impact to a specific activity such as groundwater withdrawl or contamination. It is clear that the assessment endpoints were developed with bias towards surface water receptors. Groundwater is apparently not regarded as a receptor in its own right, which may have satisfied past regulatory needs but this may be increasingly challenged in the future.	All All	
1.4 Collect data and compare data that satisfies the monitoring required by regulatory approvals of oil sands developments & **1.5 Collect data that satisfies the monitoring requirements of company-specific community agreements with associated funding** For the purpose of this review it is assumed that all cooperative and/or legal binding agreements are being fulfilled by RAMP and the operators. The only caution is that changing regulations will require constant updates in montoring practices.	All	
1.6 Recognize and incorporate traditional knowledge into monitoring and assessment activities No traditional knowledge resources have been provided to the review	All	

panel to evaluate this objective. This is considered beyond the scope of the current review.		
1.7 Communicate monitoring and assessment activities, results and recommendations to communities in the Regional Municipality of Wood Buffalo, regulatory bodies and other interested parties No evaluation was made of the effectiveness of communication beyond the review of the above noted reports and online materials. The reports and resources are generally satisfactory in conveying the monitoring results with the exception of issues noted specifically in the review. The reports are generally well-written and comprehensive.	All	
1.8 Continuously review and adjust the program to incorporate monitoring results, technological advances and community concerns, and new or changed approval conditions Recommendations are noted by objective in this table and are summarized together at the end of this review.	All	
1.9 Conduct periodic peer review of the program's objectives against its results and recommend adjustments necessary for the program's success. This activity.	All	

The specific objectives of the climate and hydrology component, taken from RAMP Technical Design and Rationale, December 2009, Section 3.4.3.2 are to:

Comments	**Component**	**Recommendation**
2.1 Provide a basis for assessing EIA predictions of hydrological changes		

See discussion in Section 1.3 above. Hydrological data are limited to streamflow and water levels in a few lakes. There is limited information available to interpret the effect of evaporation, transpiration, groundwater exchange, storage changes, and surface/groundwater interaction on the water quality data in particular.		
2.2 Facilitate the interpretation of water quality, sediment quality, benthic invertebrate community, and fish population information by placing in context current hydrological conditions relative to historical mean or extreme conditions One of the main objectives of the program is to provide information that assists in interpretation of other components. Water quality, sediment quality and other data often require flow data for interpretation. It is not currently possible to download flow data as a water quality parameter. This would be a big improvement, as well as daily and seasonal historical mean to judge whether flow is higher or lower than normal. Flow-weighted averages for water quality are most useful for calculating mass fluxes of water quality consituents for chemical balances. Addition of flow-weighted values (as part of historical data report, similar to hydrology summaries, would would be a useful addition to the dataset).	CH, All	#10
2.3 Document stream-specific baseline weather and hydrological conditions to characterize natural variability There are no climate stations south of Fort McMurray. This limits ability to document stream-specific weather conditions in the southern Athabasca region.		
2.4 Support regulatory applications and meet requirements of regulatory approvals The type and suitability of monitoring data for regulatory applications needs to be assessed on a case by case basis and was therefore considered to be beyond the scope of this review.		
2.5 Support calibration and verification of regional hydrological models that form the basis of EIAs, Operational water management plans and closure reclamation drainage designs. No details on regional hydrological models have been provided for		

review.		

The specific objectives of the acid sensitive lakes components , taken from RAMP Technical Design and Rationale, December 2009, Section 3.9.1.are (Paraphrased):

Comments	**Component**	**Recommendation**
3.1 To monitor lake water chemistry as an early-warning indicator of excessive acid deposition, where acid sensitive lakes are targeted as they are expected to show changes in their buffering capacities before soils or vegetation, and could therefore provide a clear indication that acid limits were reached. The acid sensitive lakes component has its roots in Alberta Environment's Regional Sustainable Development Strategy, and to TEEM (Terrestrial Envirmental Effects Monitoring Committee). This component is structurally different than the other RAMP components and is operated more or less independently. The ASL component was first operated as part of RAMP in 1999 with the objective of monitoring lake water chemistry as an early warning indicator of excessive acid deposition. Initially a network of 32 lakes were sampled annually in late summer but by 2002 an additional 18 lakes were added, bringing the total lakes under investigation to 50. Ten lakes were targeted for seasonal sampling beginning in 2004 to address concerns that seasonal variations were not being adequately characterized despite being potentially large. The key impact pathway identified in EIAs was from generation, atmospheric transport and deposition of acidifying emissions. The rational for choice of the 50 study lakes is not provided in the "2009 Design and Rationale" document. The reason for selection of these specific lakes needs more justification considering that the strategy deviates from standardized survey practices used in other acid sensitivity surveys which typically use a stratified-random sampling design to ensure that lakes are selected to represent natural variability within a range of size classes. There is a clear acknowledgement in the 2009 Technical report that the 50 lakes are not representative of natural	ASL	

variability, but rather that more acid sensitive systems were deliberately chosen. The report states "the percentage of RAMP ASL lakes in which the modeled PAI is greater than the critical load in 2009 (34% to 38%, Table 7.5-3) is higher than the 8% of 399 regional lakes reported in a study conducted for the NOxSOx Management Working Group within CEMA (WRS 2006). Although the use of more acid sensitive lakes is in isolation a defensible approach, the perception of the public may be that lakes are more under threat in the oil sands region near primary NOxSOx sources. And it may be beneficial to conduct an additional one-time survey using the Environment Canada approach for comparison (e.g. Jeffries et al. 2010, Gibson et al. 2010). Such a comparison would show that lakes in Boreal Plains areas in the vicinity of development are much less sensitive to acid deposition than Shield lakes in Saskatchewan and Alberta.	ASL	#11
Three concepts are not well presented in both the ASL component report and the Design and Rationale document. These concepts include the methodology for calculating Potential Acidifying Emissions at each site, and also methodologies for calculating the hydrometric and isotope mass balance estimates of critical loads.	ASL	
Review of the reports and personal communications with the component coordinator also indicate that isotope mass balance estimates of critical loads have been incorrectly reported in some cases. For example, in the Design and Rationale document, IMB results for 2002 to 2006 are attributed to Bennett et al. (2008), whereas they only reported average values for 2002 to 2004. In the 2009 technical report, IMB estimates for 2009 are misleading, and they use IMB averages for 2002-2008 (not clear). In all cases hydrometric estimates use static (constant hydrology) to run the critical loads model.The apparent insensitivity of critical loads to methodology used for estimation of hydrology, as summarized in Table 7.5-2, is inconsistent with results presented by Gibson et al. 2010, which show great sensitivity to hydrology .	ASL	#12
Gibson et al. (2010) illustrate that hydrologic variability may be a secondary driver of acidification. Reduced runoff (or water yield) by drought can potentially cumulatively accelerate acid sensitivity, and factors such as permafrost melt can enhance water yield and mitigate acid sensitivity. The problem with the latter is that permafrost melt may subside over time. Even if CL is not treated as a static parameter, PAI vs. CL should be retained as an indictor of acid sensitivity as it captures the fundamental processes driving acid sensitivity.	ASL, CH	
Unpublished $\delta^{34}S$ vs SO_4 data from the 50 RAMP lakes from 2005 reveals that higher concentrations of SO_4 are generally associated with		

progressively more negative $\delta^{34}S$, with values approaching -10 ‰ CDT. The main conclusion of this analysis is that this pattern is suggestive of pyrite/sulfide oxidation in shales as the primary source of sulfate in the 50 RAMP lakes. This also implies that acid deposition is not the primary source of accumulated SO_4 in the acid sensitive lakes.	ASL	

Other General Comments on Accessibility of Data: Online Database, Query, Maps

Comments	**Component**	**Recommendation**
Online maps were not functioning in my browser at the times I was accessing the web site for the purposes of the review, and no instructions are provided to fix such problems. The browser interface suffers from some utilitiy issues such as lack of ability to identify and select baseline and test stations, and as mentioned previously, ability to download flow data as WQ parameter. Flux weighted water quality data should be accessible for all stations.	All	#13
The online database summarizes discharge records for 2002-2008 from 90 stations, including 16 Oil Sands Operator run stations, 49 RAMP stations, and 25 Government-run stations. For water level it appears there are records available only for 2000-2002 for a total of 56 stations, 3 Oil Sands Operator-run stations, 50 RAMP stations, and 3 Government run stations. Why are data not available for all years betweem 2000-2008? For climate data there are 47 stations summarized, 20 RAMP stations and 27 Government Stations. It is not possible from this interface to distinguish between test and baseline stations and it is not obvious which stations are operational and which stations have been discontinued. An attempt was made to reconcile these numbers with the summary provided in Table 3.2, Design and Rationale, but this was not entirely successful.		
It appears that some of the stations may be missing A,B flags to indicate times of backwater due to ice. Kearl Lake outlet has no A,B flags prior to 13 january 2009. If ice-affected then flags should be shown as B. If data are not available then this should be recorded as “not available”	CH	#14

Only selected climate station data are available including daily max./min./mean temperature, total rainfall, total snowfall, total precipitation, and snow on the ground. Although relative humidity, barometric pressure, wind speed and direction, and solar radiation data are collected at some sites, these data are not available through the online data access point. Nearest neighbor or average precipitation, or interpolated values for individual gauging stations would be nice addition	CH	#15

Concluding Remarks

Overall, the RAMP program substantially meets its basic objectives at present, although there is certainly room for improvement to be made in many areas. Changes in the monitoring strategy will most likely be required in future as the degree of development increases beyond the threshold where simple reach water balance models can be utilized. Emphasis on development of improved cooperation with related organizations and networks (i.e. CEMA, GOWN, WSC, LTRN) would be a cost-effective way to significantly improve the overall accessibility of monitor ing information, and to improve perception of the public and scientific community of the comprehensiveness of environmental monitoring programs in the Oil Sands region. Development and evaluation of more comprehensive mechanistic water cycle and geochemical models might lead to better understanding of both natural variability as well as cumulative impacts in the region. Models need to incorporate processes such as surface/groundwater interaction that are particularly important for understanding of evolution of water quality in rivers in the region. Addition of new monitoring should not lag behind development in the region or there is a risk that baseline conditions may not be adequately characterized.

Summary of Recommendations

#	Component	Issue	Recommendation	Rationale
1	Climate and Hydrology	Time-series and spatial coverage of streamflow gauging for both baseline and test stations needs to be extended to provide representative records of average flows, return flows, low flows, and high flows as well as to enable statistical trend analysis	Continue time series monitoring at all possible stations	Continuation of existing stations is the obvious remedy for short period of record at many stations
2	Climate and Hydrology, Water Quality	Snow surveys are adequate for estimation of water equivaltent but might be more effective if monitored for PAHs and other pollutants such as heavy metals	Add PAH and heavy metals to routine snowpack sampling	Kelly-Schindler research suggests hat these constituents require further monitoring in snowpack
3	Climate and Hydrology, Water Quality	Groundwater monitoring is routinely carried out by operators and incorporated within programs such as AENV GOWN but is not adequately interfaced with RAMP to allow data sharing	Groundwater level and water quality data in 20-30 wells needs to be made accessible via RAMP for evaluation of storage effects and water quality modeling. Operators and or GOWN may be able to provide required data.	Groundwater has a significant impact on surface water quality and quantity on both a reach-wise and regional basis. Groundwater influence needs to be included for comprehensive evaluation of regional and cumulative effects

4	Water Quality, Climate and Hydrology, All	Discrete groundwater inputs are known to occur on some river and tributary reaches in both developed and undeveloped area and can be influential in water quality, benthos and fish habitat	EM terrain conductivity surveys should be used to map position of major seepage inputs to the tributaries and rivers. Inputs should be sampled and characterized for isotopic and geochemical characteristics to identify source formation(s) and influence on water quality.	Kelly-Schindler have suggested that contact with McMurray Formation materials along reaches should change water quality. Our research suggests that the main influence is inflow from discrete seeps. Sampling and characterization of the distribution of seeps could help explain observed water quality variations.
5	Climate and Hydrology, All	Use of naturalized flow strategy to characterize impacts	Use of more comprehensive hydrological models should be considered	As cumulative impacts increase, simple reach balance models may not be effective at capturing the causes for changes in hydrological conditions along river reaches.
6	Water Quality	Water quality analyses do not permit unique fingerprinting of sources of organic compounds	Water quality monitoring should consider using FT-IRCMS to scan for natural organic compounds as a compliment to naphthenic acid partitioning for fingerprinting natural and anthropogenic sources	Source apportionment may assist in verifying the contribution of naphthenic acids and other organic compounds of interest along various reaches of the river

7a,7b	Climate and Hydrology, Water Quality, Acid Sensitive Lakes	New baseline streamflow and lake, wetland, soil water level stations need to be added to maintain baseline/test ratio and to capture storage changes on the watershed	New baseline stations should target acid sensitive lake watersheds to provide cross-linkage and additional records for calibration of isotope mass balance	The benefits of cross-component linkage, particularly for ASL, is self-evident.
8	Climate and Hydrology, Water Quality, Acid Sensitive Lakes	Additional stations could include the Nexen lakes special project	Improved coverage in water quality, acid sensitive lakes and lake level storage can take advantage of stations already monitored under special projects	New stations that build on previous work are preferred to sites with no background data.
9	All	Qualitative criteria for impact assessment apparently use different thresholds for mining and insitu projects	Clarification is required on choice of thresholds for projects/areas	Need standardized impacts
10	Water Quality, Climate and Hydrology	Flow data are not currently downloadable as a water quality parameter, despite acknowledgement that hydrology data is collected in part to support interpretation of other data	Make flow a water quality parameter and provide flow-weighted summaries of water quality data	This will enable easier access to both water quality and water quantity parameters

11	Acid Sensitive Lakes	The ASL program, which looks at highly acid sensitive systems in a relatively insensitive region, is structured differently than other ASL programs in Canada, the latter of which seek to characterize natural variability in sensitivity. A natural variability survey would improve comparability with other programs and improve clarity and optics of RAMPs ASL	Conduct one-time random stratified sampling program to characterize natural variability in runoff (water yield) and acid sensitivity (in a group of randomly selected lakes) as comparative dataset to RAMPs ASL. The WRS (2004) survey may be sufficient for this purpose if water samples are still archived and can be run for stable isotopes of water.	Improved comparison with federal and provincial programs, increased appreciation for low sensitivity of boreal plains lakes as compared to shield lakes.
12	Acid Sensitive lakes	Poor presentation of PAI, isotope mass balance and hydrometric method comparisons	Add improved description of methodology for IMB and review of estimates bases on IMB vs hydrometric methods	Improved clarity in methodology is essential for improving credibility of ASL monitoring activities
13	All	Online maps are not functional or user friendly	Improvements in browser interface suggested	Significant accessibility improvements could be achieved through minor software upgrades/fixes
14	Climate and Hydrology	Some difficulty in locating data from hydrometric stations. And some qualifier flags appear to be missing, particularly from winter streamflow records	Check data availability for all years. Check that A,B flags are included in hydrometric records. Where appropriate, add not available to past records where no information is available	Minor improvement in online record of discharge recommended to improve archives

15	Climate and Hydrology	Only limited weather station data are made available on the web site, and coverage of climate stations is poor for the region south of Fort McMurray	Addition of a climate station south of Fort McMurray would be advantageous to allow for river-specific weather and hydrologic information to be monitored in the southern Athabasca region. More complete data records should be made available including information required to estimate evaporation and transpiration	River-specific information is required to fulfill objective 2.3 above, i.e. to "document stream-specific baseline weather conditions to characterize natural variability". This information is also important for hydrological modeling.

APPENDIX C

Dr. George Dixon Review

25 September 2010

REVIEW OF THE WATER QUALITY AND ACID SENSITIVE LAKES (ASL) COMPONENTS OF THE REGIONAL AQUATIC MONITORING PROGRAM (RAMP)

Prepared By: D. George Dixon

Introduction and General Comments

In order to complete my review of the Water Quality and ASL components of RAMP I reviewed the following documents:

1. Technical Design and Rationale, RAMP, December, 2009, including Appendices A1 to A4. (RAMP 2009a)
2. Regional Aquatics Monitoring Program 2009 Technical Report, RAMP, 2010, including Appendices. (RAMP 2010)
3. Regional Aquatics Monitoring Program 2008 Technical Report, RAMP, 2009. Only those sections pertaining to the ASL seasonal variability study. (RAMP 2009b)
4. Oil sands Regional Aquatic Monitoring Program (RAMP) Scientific Peer Review of the Five Year Report (1997-2001). Ayles et. al., 2004.
5. The RAMP websites, both public and members.

Although all of the material was reviewed, my comments have largely to do with only the Water Quality and ASL components of the work. The review was directed to provide comment on the degree to which the RAMP program was achieving three major objectives stated on pages 1-1 to 1-2 of the Technical Design and Rationale document (RAMP 2009a). Specifically:

1. Monitor aquatic environments in the oil sands region to detect and assess cumulative effects and regional trends.
2. Collect baseline data to characterize variability in the oil sands area
3. Collect and compare data against which predictions contained in Environmental Impact Assessments (EIAs) can be assessed.

The review also took into consideration, although to a lesser degree, the following two RAMP objectives:

1A. Communicate monitoring and assessment activities, results and recommendations to communities in the Regional Municipality of Wood Buffalo, regulatory agencies and other interested parties.

2A. Continuously review and adjust the program to incorporate monitoring results, technological advances and community concerns, and new or changed approval conditions.

As far as I can determine, RAMP has a two-tiered governance structure. The technical aspects of RAMP are the responsibility of the Technical Program Committee (with three sub groups) which reports to a management committee, the RAMP Steering Committee. The two groups appear to be sufficiently integrated to ensure that recommendations for change from the technical committee are indeed implemented by the steering committee. I have one concern with this structure. It appears that there is little opportunity for external input to the process on an ongoing basis. External input would appear to occur only once every five years through the current RAMP review process. Considering the rate of increase in oil sands activity in Northern Alberta, and the rapidly changing nature of the science, review once every five years may not be the most effective method to achieve external oversight. Alternatively (or additionally) RAMP could consider the establishment of a standing external review committee to comment on the activities of the Technical Program Committee. There is a second issue that this touches on: the degree to which RAMP activities are coordinated with other aquatic monitoring programs that are ongoing in the region. These appear to be increasing, and a standing external committee could also be tasked with contributing to the integration.

There has been criticism in the past with respect to the clarity and availability of the RAMP data. As such, I spent some time going through the data bases in both the Public and Members RAMP websites checking on the format, ease of accessibility and clarity of presentation of the data. I had no difficulty finding the information that I required in a reasonable timeframe and in a format consistent with good data management. In the documentation which I reviewed, however, I was unable to determine what methods are being used to assure that the data are free from clerical errors. I suggest that if a protocol for accuracy in data entry is in place it should be reported; if one is not in place it should be implemented.

While I had little difficulty obtaining the information that I needed, I have concerns that the program may suffer from lack of an overall communications strategy to make the knowledge obtained from RAMP available to external parties on a number of levels. I suggest that RAMP management clearly define who they want to inform (and about what) and develop a communications strategy to achieve their goals.

Water Quality

As discussed on page 3-2 of the Technical Design and Rationale document (RAMP 2009a), RAMP uses a combination of stressor based and effects based monitoring endpoints. In general the former are (or can be) predictive of change at the effects level, but they are variable both spatially and temporally (particularly in river systems) and identifying exposure thresholds for the onset of impacts at the effects level is rarely straight forward. Effects monitoring is less variable and more integrative through time, but the observations are retrospective (the impact has

already occurred) and identifying the causative agent or agents (with a view to remediation) is often difficult. The use of the two in combination, as undertaken by RAMP, is the best strategic approach, and significant progress in integrating the two (common stations for water, sediment and biological endpoints etc.) has been made over the last five years. Having said that, there is still progress to be made as outlined on page 3-5 of the design document (RAMP 2009a). This should be given priority.

Water quality monitoring falls firmly into the area of stressor-based monitoring, and the endpoints are notoriously variable both seasonally and temporally, particularly in river systems. The nature of this variability and the challenges in detecting real change in water quality data sets are discussed in detail starting on page 3-57 of the design document (RAMP 2009a). There are numerous analytical approaches that have been applied to data to increase the ability to detect meaningful change in water quality parameters. On balance, the approach that has been applied to the RAMP data (p. 3-57 to 3-65, RAMP 2009a) is perhaps the best for addressing this issue that we currently have. It essentially has five components: comparison to natural variability in baseline conditions (using a regional baseline approach), analysis of temporal trends (both by station and for the Athabasca as a whole), ion balance by station, comparison to water quality guidelines, and development of a water quality index. The only component of this that gives me some concern is the use of a regional baseline approach. While valid, the approach relies on a detailed and comprehensive data set on the regional baseline (reference) condition. While it is apparent that every effort has been made to identify and use all of the baseline data available, it is not clear that the baseline data set is sufficiently robust. This is addressed in greater detail below with the discussion around reference sites.

While individual water quality measurements can be highly variable (particularly in rivers), and give a “snapshot” of concentration for a very narrow window in space and time, there are analytical methods that can integrate contaminant concentrations over longer periods. RAMP has already done a preliminary assessment of one of these, semi-permeable membrane dialysis (SPMD), for the determination of polycyclic aromatic hydrocarbon (PAH) concentrations. In addition to an integrated concentration through time, SPMDs data can be used to estimate aqueous concentrations of chemical constituents (i.e. alkylated PAHs) whose concentration are consistently below detection limits in water samples. If sufficient information is available on flow rates during the time over which the SPMD is in a river system, estimates of contaminant loading can also be obtained. Estimates of loading are even more easily obtained for lake systems since aspects of hydrology are usually less variable and more easily obtained. As such, SPMDs could be used to determine atmospheric loading of PAHs in lake systems downwind for oil sands activity. While PAHs are often the target of SPMD work, it should also be possible to modify the characteristics of the SPMD membrane (and the dialysis material within the membrane) to obtain measurements for naphthenates. In addition, there are now systems available using ion exchange resins that can be used to obtain data on metal concentrations similar to that available for PAHs using SPMDs. In summary I suggest that the RAMP program

consider supplementing the current monitoring program with in situ assimilative devices where appropriate.

I reviewed the water quality variables measured by RAMP (page 3-39, RAMP 2009a). There has been some evolution of the measured parameters since the inception of RAMP; the principal justification for deleting some parameters over time has been consistent non-detect results. This is justified. As the list sits at present, it appears to me to be comprehensive and appropriate. There may be situations where other parameters will need to be added at some sites to address specific questions that arise from the larger monitoring program, but I see no reason to add any chemical parameters to the core program at this time. As an aside, I note that for the sediment quality variables (Table 3.20, RAMP 2009a), pore-water naphthenates are not included as one of the measured parameters. Since these are oil-sands chemicals of concern, what is the basis for exclusion? While I recognize that they are often considered too hydrophilic to partition to sediment, it might be appropriate to demonstrate that they are in fact not present at significant concentrations.

I also reviewed the 2009 RAMP Standard Operating Procedures (Appendix A4, RAMP 2009a) with a view to the adequacy of the sampling procedures. I find no fault with the sampling procedures, particularly with respect to QA/QC for sample handling. I was, however, unable to find reference to the actual analytical techniques used to determine the concentrations of the analytes in question. While I expect that they are contained in the Standard Operating Procedures (SOPs) of the analytical labs involved, these SOPs should be included, particularly if different labs are being used for the analytical work through time.

This is important when one is trying to determine if the methods have changed from year to year, and whether the change will impact on year-to-year comparison of the data. I will use the case of naphthenic acids (NA) by way of example. Our understanding of the chemistry of NA is evolving rapidly, as is the analytical methodology to determine the “appropriate” concentration that correlates to biological response. The use of different analytical techniques through time is thus to be expected. If a change takes place, there has to be some way of “translating” the results from the new method so that they can be compared to previous data for the purpose of trend analysis. One way of achieving this is (for the year of introduction of the new method) to complete duplicate analysis using both the old and new methods, the presumption being that the relationship between the two could be used to convert the data to a common base. I suggest that this approach, or an appropriate equivalent, be added to standard procedures as the program moves forward.

RAMP uses fall water quality sampling as an overall surrogate for annual water quality. On balance I see this as a reasonable approach, particularly for long-term trend analysis and characterization of baseline. Notwithstanding, RAMP management has recognized that the fall values may not provide an adequate representation of the risks associated with chemical concentrations at other seasons of the year (page 3-65, RAMP 2009a; I note that this section is

included in the report twice). I agree; this is particularly true for spring when systems can be subjected to increased runoff which can potentially deliver a pulse of contaminants which have accumulated (possibly through atmospheric deposition) in snow and on ice through the winter. When undertaking a risk assessment it is normal to base estimates of the potential for risk of impact on the maximum seasonal concentrations, as opposed to the minimum, mean or median values. I was particularly concerned by the statement in paragraph 2, line 5, page 3-66 (RAMP 2009a) that "guideline exceedances occurred twice as often in spring". I was unable to determine if this was the same for both baseline and test sites. I suggest that the current approach of undertaking selected studies on seasonal variability in water quality data to supplement the fall monitoring program be given increased priority. This may not, at least initially, require expanded sampling, but rather reexamination of existing data to answer the following question. Are estimates of the potential for cumulative impact on aquatic biota based on fall data the same as estimates based on data for other seasons?

At the core of the RAMP program is the comparison of chemical and biological endpoints between test (possibly impacted) and reference (baseline) sites. In river systems this uses elements of both an upstream/downstream approach as well as a parallel watershed approach. As activity has increased in the region, sites that were at one point considered to be baseline have been converted to test as they come under the influence of development. Since it would appear that no new reference sites have been added, the proportion of reference sites in the total number of sites sampled has decreased; it would appear that it will continue to decrease. One of the risks in monitoring programs is to devote more sampling effort to test sites relative to reference sites, the assumption being that the reference sites are less variable, and don't actually contribute to determining if an impact has occurred. This of course erodes the statistical power of direct comparisons. The RAMP program deals with this by using a regional baseline approach, but even if this approach is used, one must have a robust reference data set. My conclusion is that efforts should be made to assess the number of baseline sites that would be appropriate given the current number of test sites and increase the number of reference sites to that level. This is significant not only for site comparison under RAMP objectives one and three above, but speaks directly to objective two, collection of baseline data to characterize variability in the oil sands area.

The comments in the previous paragraph are largely directed to the situation in river systems. The same concerns apply to lake systems in the region, but perhaps more so and with a few additional implications. While the case can be made that RAMP objectives one through three (above) are being achieved to a reasonable degree for river systems in the region, the same cannot be said for lake systems, particularly with respect to objective two. At present four lakes are monitored within the core RAMP program, two test and two reference. Reference data on two lakes is not sufficient to "characterize variability in the oil sands area". The objective is being met to some degree for water quality by the acid sensitive lake component, but is absent for the biological components.

There is a second aspect to the issue of lake monitoring. There has been some concern expressed that atmospheric transport and deposition of chemicals of concern to lakes in the region could present a risk to the viability of those systems. As currently structured the RAMP program would be largely unable to detect such deposition or impacts. With that in mind, however, it is not clear to me that the original conceptual approach of RAMP included the either characterization of baseline variability in lakes or the possible impacts of atmospheric deposition of chemicals of concern into regional lakes.

Acid Sensitive Lakes (ASL) Component

The ASL component was added to the RAMP program in 1999 with a well defined mandate "to monitor lake water chemistry as an early-warning indicator of excessive acid deposition" (page 3-138, RAMP 2009a). While the program has evolved, it has done so in a logical fashion and, with sites and components have been added, the original core design has been maintained to allow long-term trend analysis. The number and distribution of the lakes sampled is adequate for both the monitoring function and the requirements for establishing the range of natural variability.

The list of water quality variables measured (Table 3.40, RAMP 2009a) is fully appropriate and I see no need for additions. Total and dissolved metals analysis was added in 2002, presumably to both track trends in metal concentration and deposition, and to determine if acidification was changing the relative proportions of total versus dissolved (as a surrogate for free metal) concentrations. The concentrations of free (bioavailable) metal in natural systems is now most often determined from total metal through geochemical speciation modeling, followed by biotic ligand modeling to predict environmental risk. This requires not only metal concentrations, but an understanding of the concentrations of a number of co-ions in the system. I note that the total suite of measured variables includes all of this data, should metal speciation analysis be required in the future. I am not suggesting that this is required at present; trend analysis of total and dissolved metals is adequate for the current need.

The analysis of the ASL data is a multistep process involving: 1. Between-year comparisons of endpoints over all 50 lakes (by ANOVA); 2. Calculation of critical acidity loads (using modified Henriksen steady state modeling) and comparison to Potential Acid Input values; 3. Trend analysis of endpoints in individual lakes (Mann-Kendall trend analysis); 4. Graphical trend analysis of measurement endpoints; and 5. Analysis of metal concentrations by comparison to existing regulatory guidelines. The analysis is comprehensive and appropriate to the objectives of the program.

In addition to samples for analysis of water quality variables, samples of phytoplankton and zooplankton are taken and "stored at AENV pending future analysis" (page 3-148, RAMP 2009a). While I presume that these are stored for analysis should a trend towards acidification be noted in the chemical analysis, there is no indication in the technical design and rationale

document of the actual purpose. I suggest this be included. I am not suggesting that they be fully characterized to add a biological component to monitoring at this time; the chemical monitoring is sufficient to meet the current stated monitoring goals.

Water sampling for the overall ASL program takes place once a year in late summer to early fall. As with the core RAMP program discussed above under water quality, there have been questions as to whether or not this once-a-year sampling adequately represents the state of the lakes with respect to acidification. In response to this question ten of the ASL lakes have (since 2002) been sampled seasonally (for a subset of the water quality parameters) by Alberta Environment. A summary of the results is given on pages H-13, 14 of RAMP 2009b). To summarize, it appears that the parameters can be highly variable. Once again, it is not apparent that sampling in fall gives a full understanding of the environmental risk. I suggest that the data be fully analyzed to determine the strength and weaknesses of fall sampling and that these be clearly stated. Stated another way, fall sampling has to be more fully justified.

APPENDIX D

Dr. Monique Dube Review

Pugsley/Dubé Consulting Inc.
Site 611 Box 7 RR6
Saskatoon, SK S7K 1M2

September 20, 2010

1.0 Introduction - <u>Review and Comments on Water Quality Sampling Program for RAMP</u>

I am pleased to provide below my synthesis and summary of this review as requested. I appreciate your patience in receipt of my response. The delay was due to reviewing and considering my comments in conjunction with the benthic review of Dr. Kelly Munkittrick as well as review and consideration of external publications on the oil sands submitted by other authors.

2.0 Review Approach

The overall mandate of RAMP is to: *determine, evaluate, and communicate the state of the aquatic environment and any changes that may result from cumulative resource development within the Regional Municipality of Wood Buffalo.*

The objectives of RAMP are to:

- monitor aquatic environments in the Athabasca oil sands region to detect and assess cumulative effects and regional trends;
- collect *baseline* data to characterize variability in the Athabasca oil sands region;
- collect and compare data against which predictions contained in Environmental Impact Assessments (EIAs) can be assessed;
- collect data that assists with the monitoring required by regulatory approvals of oil sands and other developments;
- collect data that assists with the monitoring requirements of company-specific community agreements with associated funding;
- recognize and incorporate traditional knowledge into monitoring and assessment activities;
- communicate monitoring and assessment activities, results and recommendations to communities in the Regional Municipality of Wood Buffalo, regulatory agencies and other interested parties;
- continuously review and adjust the program to incorporate monitoring results, technological advances and community concerns and new or changed approval conditions; and
- conduct a periodic peer review of the program's objectives against its results, and to recommend adjustments necessary for the program's success.

Monique Dubé, Ph.D.
Principle
Phone: +1 306-292-7451
Email: Monique.dube@usask.ca

The 2010 reviewers were asked to evaluate as to whether the current RAMP program is meeting the following objectives (as outlined in the 2009 Design and Rationale document):

- monitor aquatic environments in the oil sands region to detect and assess cumulative effects and regional trends;
- collect baseline data to characterize variability in the oil sands area;
- collect and compare data against which predictions contained in Environmental Impact Assessments (EIAs) can be assessed;
- continuously review and adjust the program to incorporate monitoring results, technological advances and community concerns, and new or changed approval conditions; and
- conduct a periodic peer review of the program's objectives against its results, and recommend adjustments necessary for the program's success.

More specifically, each reviewer was asked to review the component under their expertise within the scope of the whole program; in my case, the water quality component. I have submitted my comments according to the recommended sub-headings with recommendations in tabular form at the end of the document. As a member of the independent scientific peer review and one of the lead integrators from the 2004 review and assessment, I also felt the need to return to those comments and recommendations and to review progress relative to that previous assessment.

3.0 Water Quality Component Review

3.1 Objectives of RAMP Water Quality Monitoring Program

RAMP monitors water quality in order to identify human and natural factors affecting the quality of streams and lakes in the Athabasca oil sands region. The specific objectives of the Water Quality component are to:

- develop a water quality database to verify EIA predictions, support regulatory applications and to meet requirements of regulatory approvals;
- monitor potential changes in water quality that may identify chemical inputs from point and non-point sources;
- assess the suitability of waterbodies to support aquatic life; and
- provide supporting data to facilitate the interpretation of biological surveys.

3.2 Strengths of Existing Program

Since the review in 2004 there have been improvements in the design and reporting of the monitoring program including:

- Harmonization: sediment and water quality with benthos;

- Identification of test and baseline sites and measurement of change based on a comparison of these sites and relative to effects criteria;
- Statistical testing of data commenced in 2003;
- An effort to report changes by component in a more integrated manner;
- Documentation and synthesis of EIA indicators by project and impact criteria;
- An effort to quantify land use change and watershed change based on open and closed hydrological circuits and a textual description of point source discharges and releases; and
- Revision and updating of documents describing the rationale and technical design of its monitoring program (2005 and 2009 Design and Rationale Documents).

As is indicated in the benthic review by K. Munkittrick, there has also been an increase in consistency of sites, and an increase in the number of sites where there is before and after development data, which are substantial improvements to the program.

Table 1. Synthesis of station, season and parameter (e.g., PAH) consistency across years of RAMP monitoring from 1997 to 2009

Year	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Stations	5	25	18	37	36	44	45	48	49	51	52	51	53
consistent with last year		3	13	18	32	32	42	41	46	46	50	49	48
consistent last 3 years			3	11	13	28	30	39	39	45	45	48	46
consistent > 3 years				3	10	12	27	27	37	38	44	44	46
stations sampled 2 or 3 seasons/yr	4	5	4	7	8	8	5	6	16	13	13	7	7
stations sampled 4 seasons/yr	1	7	9	7	9	21	20	20	18	9	9	12	12
consistent with last year for 4 seasons/yr		1	7	6	7	9	19	17	14	7	8	8	11
consistent 3 of > years for 4 seasons/yr			1	5	6	7	9	17	14	7	6	8	8
# stations measuring PAHs in 2 or > seasons/yr	1	7	7	6	13	6	6	6	5	5	5	5	5
# stations measuring PAHs consistent with the last year any season		1	7	5	6	8	11	8	5	5	5	5	5
# stations measuring PAHs consistent 3 or > years for 2 or > seasons/yr			1	5	5	6	7	8	5	5	5	5	5

In 1997 only 5 stations were monitored and in 2009 53 stations are monitored for water quality. There has also been an increase in the consistency of monitoring at the same stations with 46 stations in 2009 having monitoring data consistent for more than 3 years. This assessment excludes the examination however of parameter and seasonal consistency (see below).

3.3 Areas of Improvement to Program

3.3.1 Recommendations from 2004 Peer Review on RAMP Water Quality Monitoring

The review of the water quality section 1997-2001 RAMP was done by M. Dubé, N. Glozier and J. Barica. General shortcomings of the program at that time included:

- Key water quality indicators were not identified;
- Inconsistency in sampling sites and parameters measured; and
- Study design was not suited to assess change.

More specifically, concerns focused on the ability of the program to characterize existing variability, detect regional trends and cumulative effects, and monitor to verify EIA predictions. There was confusion throughout on why and how to meet the objective of characterizing variability for all monitoring components including fish and benthos. Identification of spatial and seasonal patterns affecting variability was also lacking and required quantification.

Increased emphasis on the Athabasca River, the main receiver of oil sands development activities was also identified and recommended in the 2004 review. The reviewers stated:

A major conclusion in this section, stemming from comparisons of the two AENV sites on the Athabasca River (separated by > 150km), that "cumulative development in the oil sands area had not resulted in the degradation of water quality within this stretch of the river" (Pg 4-52; section 4.3.1.3) is not warranted. The single downstream site on the Athabasca River is approximately 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.

Additionally, the validity of the statement that "inclusion of the upstream of the Embarras River site near Old Fort permits potential verification of cumulative development within the basin (Pg 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, 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.

Reference to the need for tier-based decision-making triggered by change assessments was also made. Reviewers stated: *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 author's 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?*

The major gaps identified in the 2004 review were:

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 water quality 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 is not building upon well established state-of-the science 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 5 yr report.

3.3.2 Current Review of RAMP Water Quality Monitoring

Some of the gaps and deficiencies identified in 2004 have been addressed although more accurately I would state that effort has been taken in some areas and progress has been made in some areas although the gaps have not been fully addressed. Additional gaps have also appeared.

Pugsley/Dubé Consulting Inc.
Site 611 Box 7 RR6
Saskatoon, SK S7K 1M2

Selection of Stations

As was recommended in 2004, it remains very difficult to understand as a reviewer where the monitoring stations are located relative to land disturbance and existing point source discharges; every year a map is required that overlays this information. A tabular description of developmental activities (as per 2009 Technical Report) in the absence of reference to the location of all monitoring stations does not address the previous recommendations or allow for verification of where test stations are located.

Recent literature has documented that airborne deposition of contaminants including PAHs and metals due to the high temperature combustion of fossil fuels has contaminated land and water within the vicinity of oil sands development (Kelly et al., 2010). The author's state that deposition patterns of polycyclic aromatic compounds were similar and were consistent with oil sands upgraders being an atmospheric source. Deposition patterns for other priority pollutants were consistent with local sources of airborne pollutants; for distances as far as 85 km from upgrading facilities. Transport was also seasonally dependent including direct transport to the watershed via snow melt. Kelly et al., (2010) also suggest that decades of airborne deposition have likely increased contaminant concentrations in surface soils, vegetation, snow, and runoff over a broad area of boreal forest and affected baseline station characteristics.

In an attempt to verify these findings, this reviewer began to synthesize the PAH data. In the past 5 years the most consistently sampled sites and w.r.t season for PAHs were upstream of Fort McMurray in 2 seasons (spring and fall) and 4 sites in the Muskeg River. The site upstream is the government site (ATR-UFM), 1 of 4 on the Muskeg is also provincial government monitoring. In the Athabasca River, downstream of development, PAHs have only been measured at one site in 1 season in 2002-2004. That was a RAMP site. Thus, the only consistency in PAH sampling is upstream of Fort McMurray for the Athabasca River and on the Muskeg River. Thus in the absence of airborne plume delineation information relative to water quality monitoring stations and in the absence of any significant PAH monitoring, verification of the Kelly et al (2010) study is impossible. The 2009 Technical Report states that *discontinuation of PAH analysis in water from the Athabasca River mainstem occurred in 2005, due to non-detectable or very low concentrations in nearly all water samples, and ongoing, quarterly AENV sampling of PAHs at their Athabasca mainstem locations."* Table 3.2-3 shows PAH monitoring in the Athabasca River only at the upstream location. Further, detection limits for individual PAH congeners appear to be between 20-200 ng/L where as those from the Kelly et al., (2009) are <1 ng/L.

Clearly these findings need to be assessed and location of baseline stations justified and shown to be outside of aerial contamination. A review of detection limits is also required. In the absence of monitoring data for PAHs it is very clear the airborne transport of particulates is a reality and thus monitoring must be adapted to consider this source.

Monique Dubé, Ph.D.
Principle
Phone: +1 306-292-7451
Email: Monique.dube@usask.ca

It would also be very helpful to understand which of the monitoring stations are located upstream of oil sands development and the McMurray Geologic Formation (MCMF), which stations are midstream within the McMF but upstream of mining, and which stations are downstream of development and downstream of or within the McMF.

Determination of Baseline and Characterization of Variability

Now that an effects-based monitoring approach has been adopted by RAMP as recommended in 2004 it is critical that the comparison between test and baseline sites is valid to detect a change in time and/or space if a change exists. Reviewing the water quality component emphasizes the need for refining/ justifying/ and significantly improving how the background/baseline is established. As stated in the benthic review, the variability between years and between sites is so high that there is virtually no chance of detecting an impact if one existed. Review of the water quality monitoring program echoes a similar concern.

To test the hypothesis that water quality at each sampled location is within the range of natural or baseline variability a comparison is conducted of measured water quality against a range of natural variability derived from **regional analysis of baseline data** (Section 3.4.6 in Design and Rationale document). As I understand it, a multivariate procedure is conducted where water quality data from all RAMP *baseline* water quality stations from 2002 to 2009 are pooled using Objective Classification Analysis (OCA). This indicated three major groups of stations with similar water quality types (Table 3.2-5):

- Athabasca River mainstem and delta, plus Clearwater, Christina and Horse Rivers;
- Eastern tributaries, including Steepbank, Muskeg, Firebag rivers, Fort Creek and regional lakes, as well as McLean Creek; and
- Western tributaries, including Beaver River, Poplar Creek, Mackay River, Ells, Tar and Calumet Rivers, as well as Hangingstone River.

For most stations included in the cluster analysis, samples from different years clustered closely together, indicating that water quality at these stations was consistent at specific locations across years of sampling (i.e., spatial variation was more important than temporal variation in defining cluster membership). Where multiple years of data from a station fell across different clusters, data from all years for that station were placed in a single cluster that either: (i) represented the most years of data; or (ii) included other stations from the watershed within which that station was located.

Within each cluster, data from stations designated as *baseline* were pooled to develop descriptions of regional *baseline* water quality, against which RAMP data from stations designated as *test* and *baseline* were assessed.

Firstly, clustering lakes and streams together for one group of baseline stations is ecologically disastrous and is absolutely inflating the variation of the natural condition to which test sites are being compared. The fact stated that multiple years of data from a station did fall across different clusters in some cases also indicates

issues with the clustering technique. I would highly suspect that variability among watersheds with respect to water quality to be far greater than variability within a watershed. It is impossible to validate this based on the information provided. It is difficult to determine if comparing individual observations against many baseline observations collected over several years, watersheds, lotic and lentic systems and different seasons encompassed a representative range of non-inflated "normal" that is justified.

There has been little characterization or justification of baseline for reviewers to assess. A clear picture of variation locally to regionally has not been established. Baseline should be established on a local, parameter-specific (for all years of collection vs only a partial data set) basis and compared to regional baseline. Variability also requires quantification inter-annually and by season. As stated in the benthic review of Munkittrick, *The strategy of mixing data between years is inflating the variability substantially. This is especially of concern without knowledge of the ecological significance of timing within the system.*

Further, the exclusion of water quality data collected prior to 2002 because metals data from 1997 to 2001 had higher analytical detection limits than 2002 onwards if of concern is a potential loss of 6 years of data; data which are highly valuable given the inconsistency in sampling across years and seasons. It is assumed that all data collected from those years were excluded as a result of the metals detection limit changes.

The results of Kelly et al., (2010) and Squires et al., (2010) illustrate the importance of season affecting water quality in the AR watershed. The contribution of season to the inflation of regional baseline is important to quantify. This may be an issue however based on data availability. I analyzed from 1997 to 2009 by site, year, season, and consistency in year and season combinations (Table 1). Only 8 of 53 sites (15% of all monitoring sites (RAMP, industry and government) in 2009 were consistently monitored for 3 or > consistent years and for all 4 seasons. RAMP water quality stations are located throughout the RAMP FSA, from the upper Christina River to the Athabasca River downstream of development. Water quality is monitored annually each fall when water flows are generally low and the resulting assimilative capacity of a receiving waterbody is limited. New water quality stations located in waterbodies already monitored by RAMP are sampled seasonally (i.e., in winter, spring, summer and fall) in the first year to determine seasonal variation in water quality. Three years of seasonal baseline data are collected at stations established in new waterbodies and watercourses. What is not clear is if seasonal assessments are conducted three years after development. This does not appear to be the case.

Further, suggesting that 1997 is the starting point for baseline is "RAMP centric". The first mine opened in 1967 and Squires et al., (2010) clearly show that when the longer temporal record is considered over the entire Athabasca River, changes are most evident at the mouth of the river. Use of data prior to 1997 is essential to accurately determine baseline for a system as important as the Athabasca River.

With regards to use of the water quality index, it is not clear if calculations are done as station independent using regional percentiles as the objective. Further, use of the WQI to spatially compare stations requires use of the same variables and same benchmarks (deRosemond et al., 2009) and it is not clear if this was done. If not, then spatial comparisons are not valid.

Monique Dubé, Ph.D.
Principle
Phone: +1 306-292-7451
Email: Monique.dube@usask.ca

Effects Levels

Table 2.12 in the 2009 Design and Rationale Document outline the criteria used for impact assessment for oil sands projects. I reviewed those for water quality and find that the criteria used are project and water quality parameter specific. For some projects an effect is negligible at +/- 5% change, low at +/- 10%, moderate at 10 to 30%, and high at > 30% (e.g., TSS for Jackpine and CNH). For others (e.g., Conoco Phillips and Syncrude Aurora), < 1% is a low impact, 1-10% moderate and >10% high. Still for other projects (e.g., Suncor Firebag), no specific magnitude above background was given. No basis was provided for the differences in impact criteria. Further how, these relate to the chosen effect criteria in for example the 2009 RAMP Technical Report is unknown. Two comments on this: future EIAs on the oil sands should strongly consider consistent impact criteria that are consistent with the RAMP monitoring program. Otherwise the intent of the RAMP program to verify EIA predictions remains lost as was commented back in 2004. Secondly, the basis for establishing the impact criteria for effects assessment in the RAMP program different than the EIA impact criteria was not reported. Now that an effects-based program has been recognized as the path forward to measure change, the impact criteria used in EIAs and in RAMP must be 1) consistent and 2) must be tied to some level of decision or action both in future EIAs as well as for RAMP.

Cumulative Effects and the Athabasca River

The Athabasca River is the ultimate integrator of activities in the watershed and subsequently the dominant receiver of oil sands impacts if they occur. Studies external to the RAMP program do document changes in the Athabasca River temporally and spatially in water quantity, quality and the fate and distribution of contaminants of concern including PAHs (Squires et al., 2010, Kelly et al., 2010; Timoney and Lee 2009). These studies are not consistent with reports of effects from RAMP.

Two long-term monitoring stations exist on the Athabasca main stem, ATR-UFM and ATR-OF. Both of these stations are monitored by Alberta Environment and are intended to provide long-term seasonal data to examine longitudinal changes in river water quality through the RAMP study area. There are other stations along the mainstem upstream of the Steepbank and Muskeg Rivers but these stations are only sampled in the fall and since 2000. Sampling has been conducted downstream of all development ATR-DD (and bank affiliate samples) since 2002 in all 4 seasons but for standard water quality parameters (conventionals, major ions, nutrients, total & dissolved metals, recoverable hydrocarbons and naphthenic acids).

Increased emphasis on the Athabasca River, the main receiver of oil sands development activities was identified and recommended in the 2004 review and is emphasized again. Long term monitoring in the AR should not be the exclusive responsibility of the provincial government. Further, it was assumed back in 2004 that measuring a single station at the mouth of tributaries and at the mouth of the AR (separated by > 150km from its upstream baseline site) was adequate to assess cumulative effects. This same approach is forwarded in the report of 2009. A single downstream site on the Athabasca River that is 90 km downstream of current oil sands activity and affected by many confounding factors apart including the natural river continuum

(Squires et. Al., 2010) is simply not adequate. In the absence of understanding where monitoring stations are relative to development activities, which stations lie outside of or within the McMurray Geologic Formation, and how far monitoring stations are from the mouth and relative to other non-RAMP activities, tributaries, changes in surficial geology, etc., blanket use of mouth sites to assess cumulative effects requires better justification.

Triggers for Action

In the 2004 review, reference to the need for tier-based decision-making triggered by change assessments was made. Reviewers stated: *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 author's 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?*

The entire purpose of effect levels is to understand when a change has occurred outside a natural state and to tie action to those changes. Monitoring in the absence of a benchmark does not produce a change assessment. Monitoring when the benchmark of natural variation is inflated does not allow for accurate changes to be identified. Use of effect levels that do not link to actions and decisions is not management sound. As stated in the benthic review of Munkittrick, at least a two level tiered response framework is required. Exceeding the first trigger would increase the frequency or detail of monitoring for confirmation and second level trigger investigation monitoring for causal identification. Further, the relationship between water quality monitoring and other monitoring components for triggering action requires definition.

4.0 Discussion

4.1 Summary

This review of the water quality component attempted to bring continuity from the review of 2004 and highlights the following:

Since the review in 2004 there have been improvements in the design and reporting of the monitoring program including some harmonization (sediment and water quality with benthos), development of an effects-based approach to measure change, statistical testing of water quality data since 2003, an effort to report changes by component in a more integrated manner, documentation and synthesis of EIA indicators by project and impact criteria, an effort to quantify land use change and a textual description of point source discharges and releases; and revision of documents describing the rationale and technical design of its monitoring program. There has also been an increase in consistency of sites, and an increase in the number of sites where there is before and after development data.

There are also program deficiencies.

- Again, as per 2004, it is critical for a common understanding of where test sites are located and the development related activities (land change, water withdrawal, discharges) they are exposed to as well as their location relative to the McMF. Every year a map or series of maps are absolutely required that overlays this information.
- Demonstration in light of the recent Kelly et al., (2010) work, that existing baseline stations are outside of aerial contamination.
- Inconsistent sampling and detection limits for contaminants of concern such as PAHs requires explanation.
- The entire basis of an effects-based design is a defensible baseline to compare test sites to. There has been little characterization or justification of regional baseline for reviewers to assess. A clear picture of variation locally to regionally has not been established. Baseline should be established on a local, parameter-specific basis and compared to regional baseline. Variability also requires quantification inter-annually and by season.
- Extension of assessment to the true basin (see below) and consideration of pre-1997 data is strongly recommended.
    - Spatial comparisons of the WQI requires methodological clarification.
    - Impact criteria used in EIAs and in RAMP must be 1) consistent and 2) must be tied to some level of decision or action both in future EIAs as well as for RAMP.
    - Increased emphasis on the Athabasca River is required as was identified and recommended in the 2004. In the absence of understanding where monitoring stations are relative to development activities, which stations lie outside of or within the McMurray Geologic Formation, and how far monitoring stations are from the mouth and relative to other non-RAMP activities, tributaries, changes in surficial geology, etc., blanket use of mouth sites to assess cumulative effects requires better justification.
    - At least a two level tiered trigger response framework is required to link effects to action. Consolidation of effect levels and action triggers across monitoring components is required.

Monique Dubé, Ph.D.
Principle
Phone: +1 306-292-7451
Email: Monique.dube@usask.ca

4.2 Linkages and Integration with other Program Components

Harmonization within the aquatics program has improved but remains inadequate with respect to fisheries as well as with respect to linkages between water quality and quantity.

Harmonization beyond the RAMP program to include acid deposition, air deposition, terrestrial biodiversity and landscape diversity, traditional ecological knowledge, human exposure, etc. is also inadequate.

It is understood that harmonization is a massive effort of integration. However, the level of development of the oil sands and the potential for significant long term ecological and human health effects requires this level of integration.

Further, there are numerous other ongoing monitoring programs and studies of aquatic resources being conducted by government agencies, academia and industry. Individual oil sands companies, including both members and non-members of RAMP, undertake regular water quality monitoring in streams and rivers near their operations, to satisfy permit requirements. Several universities and government research continue to undertake studies in the oil sands region to better understand local aquatic resources and their response to regional development. Again, as was recommended in 2004, these activities must be integrated and reported considering the significance of the development.

4.3 Other Comments

While my focus was on water quality as a reviewer, there were several other aspects which appeared through the review that require documentation.

Explanation is required as to why some companies are within RAMP and others are not. It is not clear if participation in RAMP is optional in which case, existence of these other companies not contributing to RAMP provides a significant source of on integrated uncertainty for all other developments in the region. There were eight approved oil sands projects active in the RAMP FSA in 2009 whose operators were not members of RAMP in 2009.

RAMP receives and includes in its reports water quality data collected by RAMP, Alberta Environment and Industry. Water quality data stored in the RAMP database is only RAMP data. This division of data is archaic and limiting to the understanding of change in the basin and the ability to manage it.

In several parts of the RAMP documentation it is stated that analyses are conducted at the watershed/river basin level. RAMP includes only a component of the Athabasca basin and this spatial restriction to a portion of the basin limits the ability to assess the significance of any changes to the basin as a whole.

It is stated that the percentage of the area of watersheds with land change as of 2009 varies from less than 1% for many watersheds (MacKay,Ells, Christina, Hangingstone, Horse, and Firebag rivers), to 5% to 10% for the Upper Beaver watershed, to more than 10% for the Muskeg River, Fort Creek, Mills Creek, Tar River, Shipyard Lake, and McLean Creek watersheds, as well as the smaller Athabasca River tributaries from Fort McMurray to the confluence of the Firebag River. Transparent and forth right reporting is absolutely required. Thus provision of the actual percentage of change (rather than >10%) is recommended up front in executive summaries.

In the climate and hydrology sections data logger malfunctions and attrition is reported. The level of malfunction and "surprise attrition" is unacceptable for a program of this magnitude and significance and for the percentage of watershed change. Increasing age of equipment is a predictable consequence and should be part of the planning cycle. I understand accessibility may be an issue but again, if you can mine then you can measure. Horizon Climate Station estimated 5 months and 61 days data loss at best. Aurora Climate Station 45 days in total. Lyinimin Creek above Kearl Lake ongoing issues for 43 days. Muskeg River above Muskeg Creek 38 days. Tar River Lowland Tributary near the mouth 36 days. McClelland Lake Outlet above the Firebag River for 42 days. Pierre River near Fort MacKay, 31 days. Total in one year of roughly 296 days of data lost at one or more stations. This monitoring should be alarmed or automatic notification and a maximum allowable response time specified. These down times do not include the wildlife or human damages to monitoring stations which resulted in additional losses of data with again, long replacement times.

4.4 Recommendations

In response to the request of 2010 reviewers asked to evaluate as to whether the current RAMP program is meeting the following objectives (as outlined in the 2009 Design and Rationale document):

- Monitor aquatic environments in the oil sands region to detect and assess cumulative effects and regional trends; **Improved but still deficient.**
- Collect baseline data to characterize variability in the oil sands area; **Improved but still deficient.**
- Collect and compare data against which predictions contained in Environmental Impact Assessments (EIAs) can be assessed; **Improved but still deficient.**
- Continuously review and adjust the program to incorporate monitoring results, technological advances and community concerns, and new or changed approval conditions; **Somewhat although if more of the recommendations from 2004 were adopted, the results of this review would have been better.**
- Conduct a periodic peer review of the program's objectives against its results, and recommend adjustments necessary for the program's success. **Somewhat although discussion and adaptive management and tracking of recommendations with reviewers on an on-going and formal basis is strongly recommended.**

In response to the request of 2010 reviewers asked to review if the Water Quality component is meeting the following objectives:

- Develop a water quality database to verify EIA predictions, support regulatory applications and to meet requirements of regulatory approvals; **No**
- Monitor potential changes in water quality that may identify chemical inputs from point and non-point sources; **No**
- Assess the suitability of waterbodies to support aquatic life; **Improved but still deficient.**
- Provide supporting data to facilitate the interpretation of biological surveys. **Improved but still deficient.**

Table 2. Recommendations for Water Quality Review

Component	Issue or question	Recommended change	Rationale
Water Quality	Describe test site exposure conditions	Every year a map or series of maps are absolutely required that overlays this information	Reviewers need to verify and understand the exposure conditions to determine the adequacy of the monitoring program design
Water Quality	Describe baseline stations conditions	Describe or illustrate if baseline stations are inside or outside of the McMurray Geologic Formation. Assess contamination due to air emissions.	Baseline stations are critical. If they are exposed to oil sands naturally this information is important. If they are contaminated due to aerial deposition, then their value as a baseline station is limited.
Water Quality	Sampling in mainstem Athabasca	Increase sampling	It is the ultimate receiver and if changes are detected there, there are serious concerns
Water Quality	Variability	Reduce variability	Program cannot detect realistic changes
Water Quality	Variability	Must be shown on figures	Only way to have realistic limit on interpretability
Water Quality	Parameters: NAs and PAHs	Accelerate NA analysis methodology. Reassess PAH monitoring in light of Kelly et al (2009) findings.	PAHs and NAs are two predominant contaminate classes of concern; neither of which are being adequately or accurately quantified or measured. Thirteen years of monitoring have now been completed and development in the region exponential.
Water Quality	Calculation of baseline variability	Calculate within year, season, system type (lotic/lentic) and by	Present regional method decrease ability to detect a change. Baseline starting in 1997 is “RAMP-centric”.

		parameter before regional. Consideration of the longer temporal record; pre-1997.	
Water Quality	Water Quality Index	Spatial comparisons of the WQI requires methodological clarification.	Cannot compare spatially with different parameters and benchmarks. Clarification of method and application required.
Water Quality	Impact Criteria	Relate those in EIAs to RAMP and vice versa	Impact criteria used in EIAs and in RAMP must be 1) consistent and 2) must be tied to some level of decision or action both in future EIAs as well as for RAMP. Otherwise what is the point?
Water Quality	Cumulative Effects	Justify rationale for blanket mouth sampling stations as the watershed "cumulative effects" indicator stations	In the absence of understanding where monitoring stations are relative to development activities, which stations lie outside of or within the McMurray Geologic Formation, and how far monitoring stations are from the mouth and relative to other non-RAMP activities, tributaries, changes in surficial geology, etc., blanket use of mouth sites to assess cumulative effects requires better justification.
Water Quality	Interpretation of differences	Tier decisions	Need to know how often it is different from local reference, as well as subregional reference, as well as inter-annual variability. Significant effects can exist within the range of natural variability, and are important for detecting cumulative effects
Water Quality	Harmonization	Harmonize components	Necessary for increasing interpretability

Monique Dubé, Ph.D.
Principle
Phone: +1 306-292-7451
Email: Monique.dube@usask.ca

APPENDIX E

Dr. Kelly Munkittrick Review

September 20, 2010

Comments on Benthic Invertebrate Sampling Program for RAMP

The comments below are divided by issue, and are primarily in point form. Let me know if you require expansion or clarification on any concern, but they should be self-explanatory.

Strengths of Existing Program

1. Consistency in sample sites

	1998	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
sites	3	5	9	18	26	28	29	24	23	22	25
consistent with last year		0	5	7	14	25	28	19	19	18	18
once in last 3 years		3	5	5	15	26	29	19	22	20	19
consistent last 3 years		0	0	2	3	6	10	9	14	13	13
consistent > 3 years		0	0	2	2	2	5	5	11	12	12

There has been an increase in consistency of sites, and an increase in the number of sites where there is before and after development data, which are substantial improvements to the program.

2. Sampling frequency: annual

3. Sampling time: benthic sampling is conducted in the fall of each year to limit potential seasonal variability in composition of benthic communities. – could try and time this better, since each "fall" is different in terms of timing. It was not clear how consistent the dates are in terms of ecological timing (i.e. the same calendar dates in subsequent years won't be the same state in the river) and it is not clear if any work has been done looking at the potential impact of slight differences in this timing. Studies on how much variability occurs over the month period would be valuable.

Potential Areas of Improvement to Program

1. Study site selection

Previous criticisms have centred around issues related to consistency in sites, as well as not enough baseline sampling at sites pre-development

Reference sites: There are limited opportunities for finding reference sites in large rivers (especially in a regional design); focussing on smaller sites will (at minimum) double the number of sites available for each decrease in the order of river under study; these could also be standardized between watersheds and add another layer to the analysis (1st order vs. 1st order; 2nd order vs. 2nd order, etc.).

Exposure sites:

- There is the concern that the substrate in the mainstem is coarse sand – and only the most tolerant species may be present, and it is used as justification for activities in tributaries and higher likelihood of finding effects. In terms of cumulative effects there still needs to be some effort in the Athabasca River – ultimate receiver downstream, and if effects become detectable there, there is a problem.
- None of the study sites are on small streams
- Insufficient data at sites where development is anticipated – not enough advantage of the BACI design

2. Sampling design

The design does not permit nesting of data to look at the noise in any more detail. The lowest scale of investigation is the reach. These extend from 2-4 km and have benthic samples collected over one of those kilometres. There is little data available to look at how each replicate relates to its nearest neighbour (and if collapsing these replicates into a single point of a graph [and subsequently into the baseline range]) and how patchiness affects the analyses, including the calculation of baseline ranges.

Not enough baseline sampling pre-development – noise in the system is so high – power is <20% to detect a 50% difference. I think the only defensible way to proceed is with an EEM design of multiple reference and multiple exposed sites – the variability between years and between sites is so high that there is virtually no chance of detecting an impact with the current approach to analysis – the 95% CI for the means at a reference site overlaps 0 in 7/8 sites I looked at.

3. Sample size requirements

The sample sizes are fine (it is the variability that is the problem) – I disagree with the definition of critical effect size (see more below) used in determining sample size, but it is more conservative and a sample size should be fine. There are no real replicates within riffles; I disagree with the strategy in test sites of spreading replicates across riffles over a distance of a kilometer or more. It is not clear whether the mixing of plumes or dilution in this range is accounted for. A much more defensible strategy is to look at sites where you would expect the biggest chance of detecting an impact, and then zooming in to see if it is an ecologically relevant change (and I disagree with the apparent definition of ecological relevance used in the program), and zooming out to see how far downstream the change goes. Although there was not data to compare, a direct comparison should be made of variability, I suspect this substantially increases variability. The average coefficient of variation is quite high, compared to other data I have seen from the region.

4. Variability:

First and most significantly, all figures should plot the variability as SD or SEM. The baseline conditions are set with non-parametric analyses (5^{th} and 95^{th} percentile) presumably because the data are not normally distributed, but statistical comparisons proceed with ANOVA – that seems to be an inconsistency.

There are several strategies in examining variability that are not acceptable to me. For example, the exceedance of regional range of *baseline* variability for the selected measurement endpoints based on the mean and standard deviation, with regional range defined as ± 2*SD*, and statistically significant differences between measurement endpoints in *test* reaches/lakes as compared to *baseline* reaches/lakes.

There is enormous range between years, the 2 SD for example lakes data provided a range from 500 to >22000, and annual averages ranged from 1200 to >40000. The coefficient of variation for abundance from the lakes reference data between years is >110%; Parsons et al. 2010 (Parsons BG, Watmough SA, Dillon PJ, Somers KM. 2010. A bioassessment of lakes in the Athabasca Oil Sands Region, Alberta, using benthic macroinvertebrates. J Limnology 69: 105-117) had a mean SD for lakes in region reference 360 (115) and calculated a cv of 31%. I understand the reluctance to change methodologies during a long term program, but if the variability is so high using present methods, then the chance of detecting impacts is too low to be acceptable. At the very least, a study of variability should be an immediate priority based on existing benthic data, a comparison with literature, and supplemented if required with field monitoring next fall.

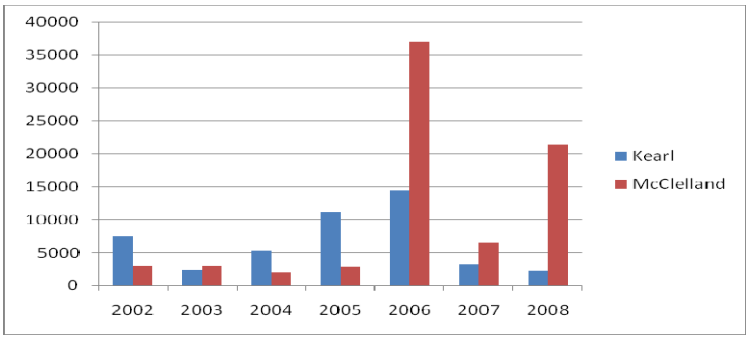


Figure 1. Baseline lake benthic abundances

The strategy of mixing data between years is inflating the variability substantially. Especially of concern without knowledge of the ecological significance of timing within the system. After this many years of sampling, there must be a better understanding of the ecological dynamics of the system and the conditions affecting benthic endpoints. In depositional areas, abundance has ranges over 60,000 org./m^2; in erosional areas the range is 40,000 org./m^2.

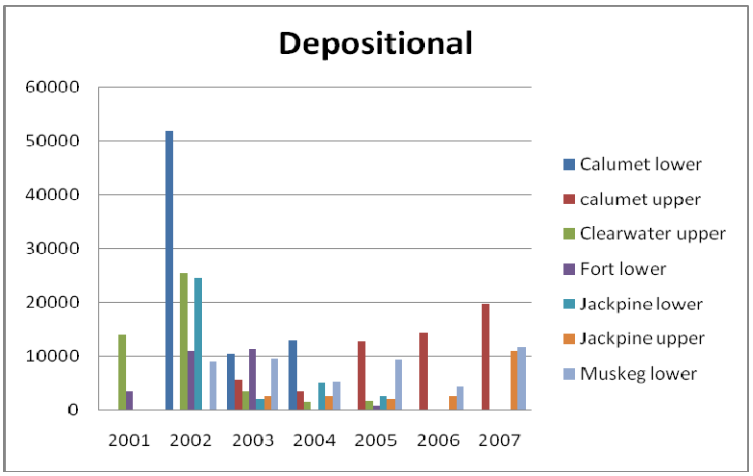


Figure 2. Baseline data for depositional areas

Regional baseline conditions were defined as the normal range of variability for measurement endpoints across all baseline sites. The normal range of variability for measurements endpoints was calculated as between the 5th percentile and 95th percentile of the measurement endpoint values. These calculations were made separately for each measurement endpoint and for each habitat. The reports needs to more explicitly specify what these ranges are and how they are changing (or if they do change) from year to year. It is not clear if the 5^{th} to 95^{th} percentile changes year to year, but it appears as if it is cumulative. It is also critical to document and report the normal range of variability for each baseline site relative to the regional baseline to better understand how individual

baseline sites are changing from year to year and contributing to increases or decreases in overall regional variability of the baseline condition.

Regional baselines compounds the noise; there is no justification for combining these sites beyond them being in the RAMP study area. It is obvious from Figure 2 that the strategy of calculating ranges across years inflates the variability and decreases the sensitivity and ability to detect a difference within a year. The within season reference site range is much less than 5^{th} to 95^{th} range across years. The variability with RAMP samples appears to be much higher than research or EEM programs, probably because of the approach to sampling. The 5^{th} to 95^{th} percentile is >> larger than 2 SD CES used in EEM, and "normal range" is 3-4 X larger than would be assumed in EEM. It is not clear what the justification is, but these decisions have substantial implications for how define changes, versus impacts versus noise. It is not clear why the current strategy has deviated from the Environmental Effects monitoring strategy of using 2 SD of the reference or baseline condition for detecting effects. The approach should have several components for assessing and reporting variability including: 1) local baseline SD versus regional SD on an annual basis; 2) the same on a seasonal basis; 3) the same on an inter-annual basis. As it is, the program lacks the sensitivity to detect responses to development that are real and are significant in terms of decision-making. It may also be possible to use the WQ clustering as a basis for developing sub-regions or analytical units; as it is there are no opportunities for grouping the sites based on specific characteristics.

Need to identify real triggers and how a range of concerns with differences based on the type of variability quantified. For example:
a) significantly different from a local reference site should trigger some change in strategy, or a warning sign
b) significantly different from regional reference data from that year should solicit confirmation – if the strategy of keeping sites spread out (which I don't support) is continued, this could include moving sites to a real near-field site
c) if it happens 2 years in a row – then an increase in number of sites
d) if it exceeds the current definition of regional interannual variability, and it is confirmed – them something needs to signify a change in monitoring – what is the cause?

5. <u>Interpretation of differences</u>

Need to fall outside of normal 3 years in a row – should tier the triggers better than this – what are the consequences to monitoring of exceedences. Would like to see at least a two level tiered response – exceeding first trigger would increase the frequency or detail of monitoring for confirmation (equivalent to extent and magnitude of EEM) and second level trigger investigation monitoring (equivalent to IOC in EEM). There are always competing

challenges of reducing noise and finding ecologically relevant changes, and this program has focused too much on using natural variability as an excuse for not detecting changes.

There is some consistency in year to year variability, but the more important aspect is that by defining normal as the range across years, they get a "baseline reference range of <1000 (5th percentile) to >40000 (95th) (Figure 3) and you can see that those values would be similar in my reduced data set (Figure 2).

Figure 3.3-2 Example of a comparison of benthic invertebrate community data against regional *baseline* data, in this case, for erosional reaches in the RAMP FSA.

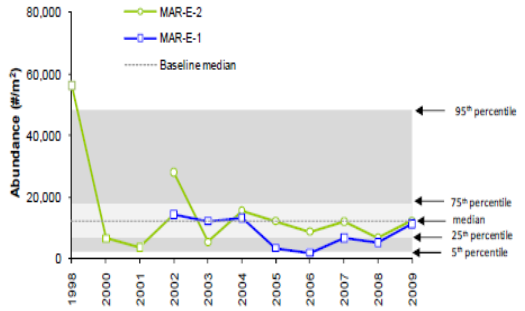


Figure 3. Example comparison from 2009 report

Normal ranges within years would be much narrower, and if defined as 2SD (which is our EEM recommendation) the average CES would be less than 1/3rd of what is being used in RAMP. This is addition to the apparent almost 3-fold increase in variability that I think they are getting from their spacing the sampling out across the reach rather than within a riffle as is normally done – so I think the program is massively inflating the variability and reducing the chance of finding an impact.

Harmonization with other studies is still not sufficient. Although benthic and sediment sampling has been harmonized since 2006, there is still not enough with other aspects. Some 2009 harmonization with fish sampling stations at the Horse River (baseline reach HOR-E-1) and the Dunkirk River (baseline reach DUR-E-1)

Discussion

There are numerous strategies and philosophies within the program that inflate variance and decrease the chances of detecting changes. Rather than looking for differences and triggering in reasonable thresholds for looking at the significance of changes, and

potential causes when they are significant, the program focuses on trying to detect serious changes, outside of natural variability. Without more analysis, I would suspect that the sensitivity is at least 5 to 9 times less sensitive (2-3x more variable, 2-3 x too large a CES) than it could be to detect change. The entire purpose of long term monitoring is to detect changes before damage is large or difficult to reverse. The program seems to go out of its way (in fish as well as benthos) to only detect changes outside of natural variability.

A program that can detect a change (even though it is not ecologically relevant) and can tell us how far downstream that change goes, and monitor over time whether it is changing, is in a much more defensible position to be a long term monitoring program. This program spreads sites out across a large area so that it will only detect changes that are outside of inter-annual variability, and has not developed or integrated triggers to increase monitoring or trigger investigations when changes exceed reasonable thresholds.

To answer the main questions posed in the review:

1) is the program design and implementation suitable to detect change in response indicators with power? No – the definition of change focuses on change that is outside natural variability. It is not clear how much of the between year variability is methodological and how much is real but it is not defensible to use it as an excuse to not find changes.

2) is there consistency across indicators in assessment and measurement of change to build a weight of evidence? No – there is not sufficient harmonization between sampling components, or with development scenarios to provide a weight-of-evidence (if you even agree that a weight-of-evidence approach is the one to use)

3) is the design suitable to examine causes of any change if changes are detected? No – it is not even sufficient to detect change.

Recommendations

See attached appendix

Appendix. Summary of Recommendations

Component	Issue or question	Recommended change	Rationale
Benthos	Pre-development baseline sites	Increase the number of sites where development is anticipated in the future	Increase time line for site-specific reference data
Benthos	Sampling in mainstem Athabasca	Increase sampling	It is the ultimate receiver and if changes are detected there, there are serious concerns
Benthos	Sampling design	Place replicates within riffles	Variability is too high, and need to study what the contribution is of spreading out the sample sites. Need to be able to detect near-field change before worry about reach-wide changes
Benthos	Variability	Reduce variability	Program can not detect realistic changes
Benthos	Variability	Must be shown on figures	Only way to have realistic limit on interpretability
Benthos	Calculation of baseline variability	Calculate within year	Present method decrease ability to detect a change by at least 3-fold
Benthos	Tier analyses	Analyze within river, and within year before regional	Strategy of EEM for need of confirmation can be adopted, but really need to increase the ability to detect reasonable change –
Benthos	Interpretation of differences	Tier decisions	Need to know how often it is different from local reference, as well as subregional reference, as well as inter-annual variability. Significant effects can exist within the range of natural variability, and are important for detecting cumulative effects
Benthos	Harmonization	Harmonize components	Necessary for increasing interpretability

APPENDIX E1

Addendum to Appendix E

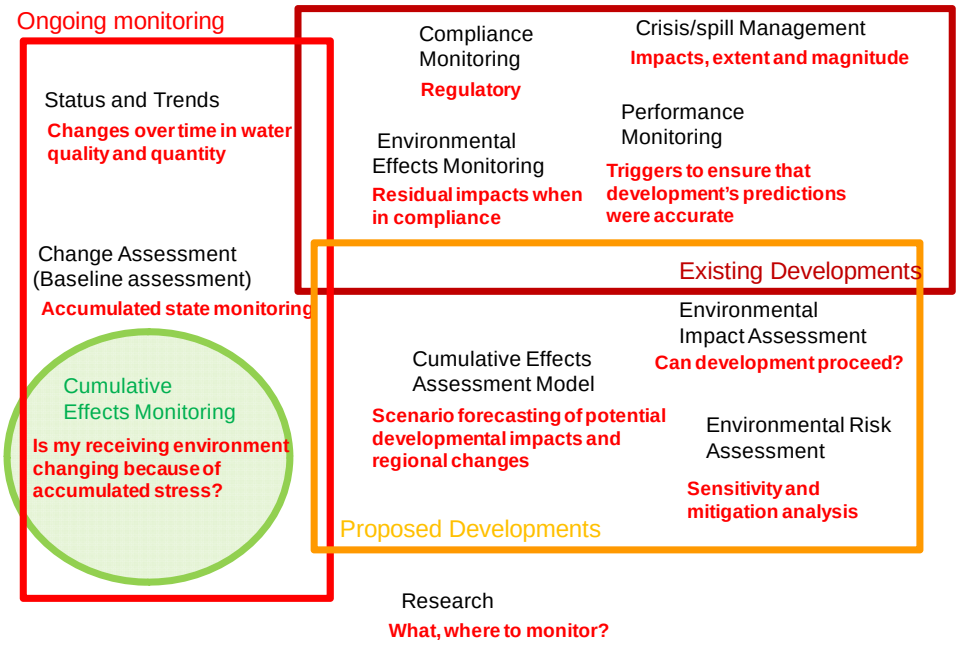
Addendum to Appendix E
Comments on Benthic Invertebrate Sampling Program for RAMP

The following comments are critical of large components of RAMP, but as I mentioned in my earlier report, there have been some positive steps forward in the program. There has been more consistency in site selection, some increases in consistency of methods, and an increase in the number of sites where there is before and after development data, which are substantial improvements to the program.

The comments below are aimed at the main discussion points from the minutes of the last meeting.

1. What type of monitoring program is RAMP?
    a. Monitoring or Surveillance? What is the purpose? – Goal? – site-specific or regional? – find stressors of find effects

This is a major issue, and needs some input from people outside of RAMP. The overall initiative needs to be tied together in a more transparent and public fashion. This program suffers from trying to be a site-specific environmental effects program (environmental effects monitoring), testing the predictions of EIA (performance monitoring), baseline monitoring for new developments, and developing a regional baseline for looking at cumulative effects. These are all important components but they have to be tied together and need a range of initiatives that are linked with similar and overlapping components.



2. How does RAMP interact with other programs in the basin? Where does it fit into the overall management framework? Who has responsibility for the various tiers within the framework?

I see the components as outlined above, with joint industry funding towards "ongoing monitoring" focusing on status and trends and baseline assessment (operating facilities and government responsibility), individual facilities responsible for both "environmental effects monitoring" and "performance monitoring" (testing EIA predictions), and proposed developments responsible for "baseline assessments" in their prospective areas. There needs to be sufficient linkages to overlap the programs (ie. so that baseline assessments are tied to indicators useful for EEM and performance, as well as are important in EIA and CEA evaluations). In some countries governments fund status and trends monitoring at regional reference sites, with the timing of sampling standardized with the industry monitoring so that both local and regional reference sites are sampled.

3. Proactive approach to study design and spatial coverage, how to best organize sampling over space and time; consider probabilistic design.

This is also a tough component, and the program has been largely in a reactive mode. Again there needs to be some sort of blending of approaches. Status and trends and regional baseline assessment can use a probabilistic approach, but EEM, performance and site-specific baseline need to have consistency in indicators. The ongoing loss of reference sites as they develop also challenges the situation, but pre-development data is essential at those sites. The long term consistency in sites and timing of sampling is essential. Ideally sites that would not be developed or are scheduled for a long time in the future would be selected and continued.

4. Integration of components
    a. processes rather than measurement endpoints -how the system works (physical, chemical, biological)
    b. Finding sources of natural variability
    c. Toxicological indicators of stress in system, which would show effects of exposure before population parameters

I am strongly in favour of developing process-oriented understanding over time. Consistency in sites and indicators and timing is critical for understanding natural variability, but I am not in favour at all of the current approach to capturing and using natural variability. I am not opposed to using regional reference data but the dramatic variability between years means that using the 5^{th} and 95^{th} percentiles constitutes a very wide range of "normal". I understand the desire to include the natural variability but it may obscure situations where effects only happen in low flow years, or in colder years, or ..... Comparing within year should be a first tier of comparison, and there needs to be the development of a real tiered and adaptive management approach. Reference sites should be compared and may be contrasted, or grouped depending on whether they co-vary (or not).

The current approach seems to make a statement that it is really variable and therefore we will only look for really big changes. Some people need to gain more confidence that significant changes will be detected. Taking another approach does mean that some time and money will be invested in chasing things that might not be real. That is why we have a confirmation cycle triggered when you know that something is outside of "normal" for this one year is a first tier of response.

Apparently CEMA has been working on triggers and thresholds, but it is not clear to me why CEMA, RAMP and EEM are not tightly linked in philosophy, operation, sampling and reporting. The conceptual model we are currently working on is to have tiers and triggers, for example:

Tier	Trigger	Question	Frequency
Basic		Are there changes	Regular
Confirmation	Statistical difference beyond a critical threshold	Can we confirm them	More often
Extent	Confirmation of changes	What is the extent and magnitude of the change	More stations
Cause	Change across a sufficient area or of a sufficient magnitude, or is getting worse	What is the cause, and if it needs to be fixed, what is the solution?	Research-oriented

There needs to be some philosophical decisions made about what constitutes a change that would be sufficient to alter development decisions, and then to develop specific predictors and indicators of those critical endpoints. As with the current program, there would need to be components of interest to local stakeholders as well as of interest to regulatory decision-making.

5. Philosophical basis for determining an effect – the selection of approach has implications for the program and its ability to ID an effect (i.e., issue of variability).

A better understanding of variability, and the size of a difference that would trigger a change in monitoring, or trigger some discussion of larger issues, needs to be developed.

6. Confirming methodological approaches, confirmation of identifications (taxonomy), which could require having more raw data available.

Needs no real explanation.

7. Inclusion of scientific advisory board, and more public availability of data. – agree and it is just about accountability and transparency.

8. Other issues - Trying to design monitoring programs is a huge challenge, and takes a lot of investment in time, in meetings, in discussion, and in philosophical convergence. There is always another damn academic "expert" who wants something different, wants a new approach, wants to use other techniques, designs, etc, and I am no different. The important thing is that a broad group of stakeholders needs to develop consensus on a core program that doesn't change.

All monitoring decisions are a compromise between competing desires for indicators with high ecological relevance, short response times, that can be linked with a cause and can be easily corrected if impairment occurs. Monitoring at the community level accepts that the endpoints are highly ecologically relevant, but that there will be a long time lag, little ability to determine a cause and challenges in reversing the impacts, and that a lot of reversible significant changes at lower levels could have been detected. Monitoring at the biochemical level compromises ecological relevance for ability to determine cause and reversible, short time lags.

EEM for metal mining took 7 years to get a consensus program. This program does need to have some more peer review, but it will need to be a significant investment of time and resources – having 3 people or 6 people to give their opinion means that there will not be broad acceptance, and the program will continue to fluctuate as more cooks put their spoon in.

Fish community monitoring in most parts of Canada is a challenge, and there are real complications in trying to use them for decision-making. To do it properly requires multiple gear, multiple seasons, and at most sites, multiple times of day. Designing a proper program requires a large investment, unless you satisfy yourself that shallow water, riffle habitat, daytime, backpack electroshocking with 3 people in September (for example) gives you the most useful and responsive information.

Population monitoring in fish, even with a huge investment, will never detect changes less than an order of magnitude in size – they just don't respond in a way that offers much to a program. But both are important for understanding relevance of lower level changes. But fish fences are an extremely expensive, highly variable, prone to failure ways to get information that has a lot of natural variability and will require decades to start to get an understanding of.

Biochemical monitoring does provide "early warning" capability, and that is an important part of "performance monitoring" which serves to monitor EIA predictions. But it is much more difficult to use for status and trends monitoring without a lot of development.

Teams of multi-stakeholders developed the EEM approaches, and while there may be some philosophical differences when dealing with non-point disruptions such as landscape clearing, the program endpoints will work for status and trends, and change assessment.

Regardless of what I think, the program needs to have a broader discussion of where things fit, how they link, what the objective is, and what the endpoints will be. Additional components can be added for specific questions or as the program evolves, but the core must be sustained to give the database necessary for cumulative effects assessment.

APPENDIX F

Dr. Joseph Flotemersch Review

RAMP Regional Aquatics Monitoring Program
Comments from review of:
Technical Design and Rationale- December 2009

Section 3.6, 1st paragraph – "sediment-quality component was folded into the benthic-invertebrate component"

Excellent that the collection of benthic macroinvertebrates now coincides with the collection of data for the sediment-quality component. This will allow investigators to better account for differences due to natural variability and improve ability to detect differences due to anthropogenic impacts.

Section 3.6.1, 2nd paragraph – "In 1998, the focus shifted from the mainstem to tributaries. The mainstem was taken out of the program for two reasons. One, benthic invertebrates in the shifting sands of the Athabasca River are typically tolerant to disturbance. In the diluted environment of the Athabasca River, it could be anticipated that the benthic fauna of the Athabasca River might not be an adequate indicator of possible changing conditions due to oil sands operations. Second, tributary rivers, typically with more stable substrates, tend to contain more sensitive benthic invertebrate taxa that are anticipated to respond well in advance of benthos from the mainstem, to oil sands development-related stressors."

This comment caught my attention, probably because much of my career has been focused on the development of methods for assessing mainstem rivers. I understand the logic here for dropping mainstem sampling but taking this position does seem to limit the ability to detect the cumulative impact of multiple tribs on the mainstem. In reviewing the methods, it seems that the only methods that have been used for sampling are the Ekman grab and the Neill-Hess Sampler. These methods have been found inadequate in many larger systems as they only sample those substrates habitats and substrates where they can be used. In most mainstem rivers, this does not include the riparian section of the river. The argument can be made that the riparian zones do not constitute the majority of the habitat, but they may and often do contain those most useful in detecting potential impacts. I would consider mainstem riparian zone sampling with D-ring or kick-nets of some type. To be effective and not clog, the mesh size used here might need to be increased from what is currently used to 500 or 600 micron.

I'd also like to comment on the statement that "benthic invertebrates in the shifting sands of the Athabasca River are typically tolerant to disturbance". Before I say much on this I think it would be best if I discussed this with Bruce in greater detail. Is this possibly due to the benthic organisms that are being sampled by the methods being used? I do understand the Athabasca River is a rather harsh environment seasonally and it would make sense that the organisms are tolerant. Still, something has to exist on that edge of tolerance; and once identified, those organism could serve as key early-warning indicators of impact. It's the nature of nature.

Section 3.6.1, 5th Paragraph – "The tributary monitoring approach adopted by RAMP has focused on the lower reach of each river to allow detection of the cumulative effects of all developments within each basin"

This identified focus would seem to support the value of mainstem sampling (which has been dropped).

Section 3.6.1 – General Comment – This section and much of the document would be much strengthened by the inclusion of citations to back up the reasoning provided by the document. Otherwise, the decisions made seem more like opinions rather than decisions well founded in the peer-reviewed literature. For example, in the last paragraph of this section it states "Benthic sampling is conducted in the fall of each year to limit potential season associated variability in composition of the benthic community". Many citations are available to justify this position.

Section 3.6.3.2, 1st Paragraph – The sampling method used as I understand it only samples riffle habitat. Consequently, a statement can only be made about the condition of the riffle habitat and its relevance as an indicator of overall system condition/health/change. Thus, the first bullet here should be modified at read

> "Collect scientifically defensible baseline and historical data to characterize variability of indices of composition of benthic invertebrate communities ***of select habitats*** in the oil sands area ***that have been shown in the literature to be indicative of overall system condition***;

The same would apply to the remaining bullets on this page (3-75) and 3-76.

Section 3.6.4 – I did look at Appendix B and did not disagree with the level of taxonomic resolution used for this study. However, I was very surprised that little diagnostic value was found in the mainstem sample given that chironomids where taken down to the genus/species level. This may again be due to the sampling method used. Additional input by the researchers is needed here to better identify why the diagnostic value is so low. Doesn't make sense to me.

Section 3.6.4, Paragraph 4 – Again, sections like this would really be strengthened by the inclusion of citations to back up what is being stated; especially since most of what is being stated should be easy to find citations for.

Section 3.6.4, General Comment – The design here does have strong statistical power, but the methods being used to collect the data may be limiting the ability to "Identify effects before they become irreversible". Ekman grab/Neill-Hess samplers are excellent quantitative sampling devices, but used alone they may not fully support the RAMP program in meeting its objectives to the extent possible. At a minimum, I would suggest some pilot samples be collected using alternative methods. Many options are available here and I would be more then willing to provide materials to review.

In the overview provide by Bruce Kilgour on July 15th, it was my understanding that riffle habitats were samples. At the first riffle, a single sample was collected, with the next 9 riffle being sampled for a total of 10 individual samples that were composited in to a single sample. This will be effective at detecting changes that impact riffle habitat, and likely will serve as a suitable indicator for overall impact. However, it will not detect a reduction in riffle habitat (identified as most productive). If changes in habitat are of a concern, then I would suggest that a multi-habitat method be considered. Habitats would be sampled in proportion to their presence in the reach. It all depends on the questions being asked. Do you want to know the condition of the riffle habitat as an indicator of overall condition, or the condition of the river overall. Both options have their pros and cons. I

suspect the program has been criticized for changing methods over the years. Therefore there is likelihood that there will be resistance to changing methods. If this is the case, one option may be to strengthen the measure of physical habitat condition.

Section 3.6.5, 1st Paragraph on page 3-79 – "Within Reaches, samples are collected from either erosional or depositional habitats, depending on which is the dominant habitat type within the tributary." Bruce seemed to indicate that they sampled riffles. Which is it? Regardless, I think it is critical that the methods being currently used be better documented to support the collection of data through time as field crews change. All too often we analyze data, find a significant change in condition, or significant difference, and can trace it back to a change in crew member, different contractors, or a change in brand of gear (although the specifications are the same). In short, document the methods more completely. What are the specs of the gear? For the Neill-Hess sampler, how is the sample collected? How deep is the substrate disturbed? What is the extent of field sampling? What is the total area sample by different gears? Is this area comparable and consistent? This information is very critical for monitoring efforts over time; such as the one that is being conducted.

Section 3.6.5.1 – Either in this section or the previous section that discusses the same material, add more specifics about the sampling methods to assure continuity through time.

As for the laboratory processing of benthic macs, a good job has been done here to detail the methods but a more details should be included that would facilitate replication of the methods by a different lab. Additional details would also help reviewers identify areas for improvement. I will try to find an example and forward this material.

RAMP Regional Aquatics Monitoring Program
Comments from review of: 2009 Technical Report, Final

Section 1.4.5.3, 1st Paragraph, 4th Bullet – Provide citations for "known tolerances of benthic taxa". Without more information, it is difficult to comment on what is being used. This information varies by author. To better assist the program, reviewers need to know what information sources are being used. In river studies, people often use the information about species that was derived from stream research. In many cases, this information does not translate well to rivers.

Section 1.4.5.3, 4th Paragraph – See comments on from review of Technical Design and Rationale document, Section 3.6.1.

Section 1.4.5.3, 5th Paragraph – "A reach consists of relatively homogeneous stretches of river ranging from 2 to 5 km in length, depending on habitat availability".

It is nice that a general length for a reach is identified but I think a more scientifically defensible method for characterizing a reach would benefits in teasing out differences due to river-type (reach-type) and thus aid in the detection of changes in the condition of the system. I would highly recommend that the system be delineated in to Functional Process Zones (FPZs)(Thorp et al. 2006, 2008), and then use these zones to account for more of the natural variability in the system. Also check out the paper at (http://onlinelibrary.wiley.com/doi/10.1002/rra.1367/pdf) for a through discussion of reach length in rivers. I can provide much more information on this approach if action is taken to move in this direction.

Thorp JH, Thoms MC, Delong MD. 2006. The riverine ecosystem syn thesis: Biocomplexity in river networks across space and time. River Research and Applications 22: 123–147.

———. 2008. The Riverine Ecosystem Synthesis. Academic Press.

General Comments from July 15th overview not included above:

Since baseline areas are slowly disappearing, consider the development of a predictive model. This would maximize the use of existing baseline data by providing a benchmark to measure condition against in the absence of a suitable baseline site in a given tributary. The process of predictive model development has become increasingly available with the development of programs to aid in their development. Consult Flotemersch et al., (submitted; available upon request) for programs available online for free.

Several times during the overview, the investigators (e.g., water quality, benthic invertebrates) mentioned the problem of suitable reference rivers and lakes. Are their reference rivers or lakes available outside the study area? If so, why are they not considered? If they cannot be used, this further strengthens the argument for development of predictive models to assess condition of resources being monitored in the study area.

During the water quality overview, the investigator mentioned that they sample both east and west banks of the river as it does not mix well. If this is true, then sampling for benthic macroinvertebrates should fully consider this detected difference in their sampling design. As is, there is no evidence that they do.

In general, the various monitoring elements of the monitoring program are not well integrated and thus not mutually supportive. The exception to this is the Benthos and sediment components. Integration and coordination of various elements is usually a rather arduous task and may result in less data being collected overall, but the larger team will be able to say much more with the data that is available. Advantages of more coordinated efforts include increased efficiency in the field and an increased diagnostic ability when changes are observed.

Many of the elements of the monitoring program appear to have different general scientific objective. For example, the water quality program has a *monitoring* focus, the fish tissue program is focused on *exposure*, and the benthic invertebrate program is focused on *effects*. I think this can be spun as either a problem of the program or a weakness. I think the best approach is to acknowledge this and then work to identify the advantages and disadvantages so this information can be used to identify a path forward.

APPENDIX F1

Addendum to Appendix F

Addendum to Appendix F
RAMP Regional Aquatics Monitoring Program

MEMORANDUM

Discussions with Joe Flotemersch of November 22

RE: RAMP Second-Order Recommendations

During discussions with Dr. Flotemersch on November 22, 2010 the following issues were put forth for discussion at the December 13th, 2010 RAMP Reviewers Meeting.

1) Surveillance versus Monitoring Program

There is a need to define the type of sampling program that RAMP is conducting. Until this is determined, it is difficult to discuss the requirements for the program. The first requirement is to determine the purpose of the program in this context.

2) Understanding Methods Used and Performance of Field Methods in Line with Data Quality Objectives

There is a need for a better understanding of performance of methods used. If biases with the methods used are expected to be in the 30 to 40 % range, then greater than 40% change would be required before there would be a discernable response. There needs to be a decision as to whether this is acceptable for the information the program needs to determine. Good templates presently exist for evaluating the variability within a sampling method.

3) Is Annual Reporting of Every Component Appropriate?

Dr. Flotemersch indicated that annual reporting of all components every year is excessive based on the expected change. He contends that there are not enough resources/scientists to accomplish this and advance the program. He indicates that there is a possibility to report on a different component each year (i.e. 4 year rotation) with a review conducted every 5th year. Reducing the need to drive annual reporting he contends will allow for more thorough study of the individual components and better science overall. He also indicates that the monies may be wasted for the amount of change expected over a one year period.

4) Use of Sensitive Species as Indicators.

The Athabasca River Basin is a harsh environment naturally. The organisms and fish that live there are by nature very hardy and tolerant. There is a challenge in asking science to find sensitive species among tolerant species.

APPENDIX G

Dr. John Post Review

An Assessment of the Regional Aquatic Monitoring Program (RAMP) with an Emphasis on the Fish Components

Prepared by

John R. Post
Department of Biological Sciences
University of Calgary
Calgary, Alberta T2N 1N4

Table of Contents

Overview

We have been asked to review the current design and progress of the Regional Aquatic Monitoring Program to address the three primary objectives in the Terms of Reference as presented in the Technical Design and Rational Report.

1. Monitor to detect and assess cumulative effects and regional trends.
2. Collect baseline data to characterize variability in the oil sands area.
3. Collect and compare data against which predictions contained in the EIAs can be assessed.

I have focussed primarily on the fish components of the program plus other program features that are useful in determining oil sands development impacts that assist in understanding impacts on individuals, populations and communities of fish in the oil sands area. The primary documents for review include the Technical Design and Rational Report (2000) and the Technical Report (2009), but also included previous Technical Reports, program reviews and data available on the RAMP web site.

The first objective involves development, testing and use of indicators to detect impacts of oil sands operations within the region. This assessment will focus on the effectiveness of fish oriented indicators to detect effects, if they occur, but since the problem of cumulative impact is really an ecosystem oriented issue, other related RAMP components were also assessed. The second objective involves the description of a baseline of fish communities in the region and spatial and temporal variation in that baseline, from which future data collection and analyses can be used to identify potential impacts. The third objective is to collect the data necessary to test site specific EIA predictions including disturbance or loss of fish habitat, changes in fish health and changes in diversity of habitat or fishes.

The fisheries component of the RAMP program has been operating since 1997 and also includes access to some data collected previous to the program initiation. As expected, the program has developed substantially over that time and includes some components that persist throughout and others that were useful in scoping the program, but no longer continue.

It is understood that these are not easy systems in which to work and that the region is one in which we have a poor understanding of much basic biology of the fish species present and their use of the main stem rivers, tributaries and lakes in the oil sands area (I think that this observation applies in general to boreal ecosystems). In this regard, the RAMP fish program has made important inroads in our basic understanding of: (a) reproductive migrations of several species in the Athabasca River and tributaries, (b) size-structure, growth and survival of several species and populations, (c) species diversity in Athabasca tributaries, (d) metal and PAH concentrations in commercial/recreational/subsistence fish species, (e) an initial attempt to develop a regionally appropriate fish community based metric of ecosystem integrity, and (f) how some of these metrics vary spatially and temporally. These have all been important components of the development of an integrated assessment of impacts of oil sands development on fish in aquatic systems in the Athabasca River basin.

The primary purpose of my review is to assess the ability of the fish program, as it exists currently, to address the 3 objectives above and to recommend approaches going forward. The review will be structured according to the existing components of the program, general recommendations and an assessment of the ability of the RAMP Program to meet the stated objectives.

Detailed Assessment of Fish Program Components

Fish Inventories

E-fishing inventories were completed in the main stem Athabasca River (10 reaches) and Clearwater River (3 reaches) in spring, summer and fall 2009. The correspondence analysis is useful for examining years that stand out from others and I presume that if temporal trends were to emerge that time could be added to the analysis. It is not clear if the analysis is done using cpue by species or percent by species. I think that the former is more useful because it can extract trends in both abundance and species composition and potentially identify differing responses to development among species. This needs to be clarified in the methods and

discussed in the results section. It is also not clear why the analysis focussed on only the KIR species rather than the whole community. This needs to be discussed and rationalized.

The length frequency analysis is extensive but rather descriptive and does not lead to useful inferences about changes over space or time. It is unclear what types of patterns in length frequency would be useful in testing hypotheses about development impacts. On the other hand, a demographic analysis of age-frequency yields growth, survival and recruitment rates which are processes that could be logically linked to hypotheses of development impacts. This is a common theme throughout the report. There is extensive statistical analysis of patterns but only minor attempts to extract useful biological insight from the data. The approach that I will argue strongly for, here and elsewhere, is an assessment of rates and processes in relation to development rather than statistical assessment of patterns which often lead to little in the way of insight into the biology of the fish or potential development impacts on that biology. Correlations can always be criticized based on lack of cause-and-effect whereas demonstration of processes lends substantial credibility to inferences.

The assessment of “recruitment” using ratios of abundance of fish below and above a biologically arbitrary size is very crude...maybe better than nothing, but age-frequency data could provide much more precise estimates. The problem is that the measure of recruits includes multiple cohorts that are summed into a single measure. The age-structure approach opens the door to a variety of recruit-stock approaches common in fisheries and provides a framework for assessing impacts.

The analysis of condition is hard to interpret from an impacts standpoint. Clearly severe negative deviations from the norm are worth pursuing. This normally results from a short term reduction in prey availability for individuals that have grown to a large size in better times and now can’t acquire sufficient prey to meet the metabolic demands for that large size. I suppose that there may also be toxicological explanations for this observation. In general, a more diagnostic measure is size-at-age and growth rate which are products for the demographic analysis discussed above. But even growth is tricky to interpret in terms of impacts because if the impact is on abundance, growth can actually increase as the impact increases due to

density-dependent growth. It really needs to be interpreted in a whole ecosystem context including info on density, prey abundance and flow regime which alters metabolic rates. Again, not simply a pattern analysis, but an assessment of biological processes is necessary to gain insight into potential development impacts.

The Athabasca River correspondence analysis has data from the 1980's. Can this also be included in the time series plots to broaden the time horizon and as a baseline pre-development?

In the Regional Synthesis section time series of species aggregated is shown (Fig. 7.4-1). Such aggregated measures are only useful if it is clear that all species have similar time trends. I would rather see time series of species individually.

Spawning Assessments

The spawning assessments in 2009 involved running a fish fence on the Muskeg River with contrasts to sampling in two earlier years. The Muskeg River is an important site for collecting this type of data since it is the tributary with the largest proportional oil sands development of the Athabasca River tributaries. Although attempted in other years, only these three years had sufficiently low spring flows to maintain the fence. The primary outcome from this work, as presented in the report is a description of the timing, species composition, run size and size-structure of the large-bodied mature fish captured during their spring reproductive migrations.

No criteria exist for assessing these observations, but I presume that the intent has been to develop a time series of reproductive effort by these species. It is not clear how this data is interpreted in regards to development impacts. These species likely spend most of the year in the main stem Athabasca River (or lake) and use the Muskeg River for spawning and maybe rearing. It is unclear how the authors envisage interpretation of the data in the context of the spatial location of the potential impacts? The 2009 data show a decrease in the abundance of some taxa with a substantial increase of suckers. It is unclear if this indicates a change in habitat

quality/quantity favouring some and detrimental to others. More data is necessary to determine this, but these are certainly large magnitude changes that warrant further exploration.

An area that has not been explored from the Muskeg fence data is the age-structure of the mature fish. This can provide estimates of rates of survival, growth, recruitment and could identify incidences of recruitment failure, if they exist. These estimates could be very useful in understanding trends in abundance and identify key processes in these populations.

The report states that "based on intermittent operation of fish fence programs...any changes related to oil sands development remain undetectable" Tech. Report 2009 p 5-84. I disagree, such data on reproductive effort/success is key to understanding population impacts. It will require more data, but represents potentially the most important of the cumulative impacts in the Athabasca watershed.

The report presents mean ages of migrants-mean age is very insensitive to variation in age structure. Full age-frequency distributions should be presented.

Since the Muskeg River is the most heavily developed tributary and also obviously provides important spawning/rearing habitat for Athabasca fish – it should explored further to assess spawning habitat, egg survival, fry survival, rearing habitat and toxicological assessments on early life history stages.

I noticed a brief statement in the overall synthesis that the Muskeg River fish fence program would not be continued after 2009. I would recommend that not only should it be continued in all years where the flow rates allowed it, but that similar programs should be implemented and maintained in all trap-able tributaries. This is not a small undertaking but will yield the best data on spawning effort, population abundance and demography of the large-bodied, spring spawning and migratory fishes. Other techniques do not yield abundance of these fishes. An alternative, suggested by Bill Franzin is to sample with a mobile gear. This will yield size- and age-frequency samples but not abundance. Spawning effort among tributaries and over years, and the resulting demographic data, would provide an assessment of the degree of success of

the large-bodied fishes in the watershed as development proceeds over the years and decades. This is neither easy nor cheap, but would provide data to address all three of the primary RAMP objectives. Unfortunately, the current extent of the data collection, lack of demographic analyses and apparent abandonment of the program provides little to meet the objectives.

Sentinel Species

The sentinel species program in 2009 involved 3 test sites in the Muskeg and Steepbank rivers and 2 baseline sites in the Horse and Dunkirk rivers all of which involve slimy sculpin. As I have recommended elsewhere I think that the most effective presentation would be thematic as was done for the sentinel species program (hidden within the Steepbank River section) rather than a watershed by watershed structure. In addition, discussion of the whole sentinel species program over years should be included in the annual reports as a living document to facilitate extracting insights developing from the program rather than the single year sampling. I will raise this issue further in the General Recommendations section.

A key requirement of the baseline:test sites approach in the Sentinel Species Program (and most of the other components of RAMP) is that there is a point source of impact. If all potential impacts are of this sort then the assessment should be robust. In instead areal transport and deposition are important, or if the test organisms can move upstream, then upstream sites may not be adequate control sites. If the case then the baseline:test contrasts will not provide robust assessment of impacts. This would result in incorrectly accepting the hypothesis of no effect when there is one! This is the same problem as low power but is more insidious in that it can't be remedied with simply more sites. This potential serious criticism of the RAMP design must be remedied by much broader spatial sampling within the watershed.

What I like about this program is that it attempts to measure processes rather than just describing patterns but I think it can be augmented to be more informative. There appears to be a habitat bias in the test:baseline contrasts with all tests as riffle and all tests as runs and clearly these two habitats have different characteristics leading to different sculpin populations.

Temperature was also measured at these sites which proved to be very useful. Benthos was not measured which may also been useful in contrasting differential growth and survival among sites. The 2009 analysis is an excellent example of the importance of ancillary physical and biological data in interpretation of patterns and developing the insight necessary to meet the program objectives.

A powerful approach involves development of habitat models for slimy sculpin from data collected over many baseline sites involving physical, chemical and biological data and then asking if there are deviations from these relationships observed in test sites. The key question is what ecosystem features do sculpin populations need to be successful and how these are quantitatively related to abundance and structure. This might sound “academic” but in fact this approach will likely yield much more useful insights into potential development impacts and mitigation approaches than a black-box “are the sites different or not”. Other information that would be useful would be reproductive measures of success such as reproductive investment and success of juveniles in baseline and test environments.

Enhanced coordination of components of RAMP with hydrology, chemistry, benthos, substrate components would enhance the interpretation of the sentinel fish program. In addition, as argued effectively by Whittier and Hughes (2008) many more sites distributed in a random but stratified design would substantially improve the match of the program to its key objectives.

The lethal sampling needs to be reinstated. The value of the reproductive investment data is crucial to assessments of potential impacts. What are the densities of sculpin sampled per km (and likely at best 40% are captured with the most efficient e-fishing program)? What would be the impact of sampling 50-100 from each reach?

There were differences in condition identified among sites. How is this interpreted? See earlier comments about condition and what it might be reflective of.

The survival of young-of-year seems to be estimated by yoy:older in the summer divided by yoy:older in the fall. Does this not reflect as much about the fate of older fish as yoy?
Should not the survival estimates of yoy be calculated as cpue of yoy/cpue yoy in fall and older

fish survival as cpue of larger/cpue larger in fall? Same thing with the contrasts to the 2006 data – use cpue2006:cpue2009?

The temperature dependent growth analysis is good and might be even more predictive if benthos abundance was included in the model – another reason for development of mechanistic models and then asking if measures like growth, survival and recruitment deviate from these models.

Tissue Analysis

The tissue analysis work serves two purposes, one directly related to RAMP in assessing contaminants in fishes relating to oil sands development and the second related more to regional contaminants from a public health perspective. The data appears spotty with Muskeg River [Hg] showing substantial increases since 1975 but apparently no data available for the Steepbank and Muskeg rivers since that time. The plot also showed no Athabasca River [Hg] data since 1992? It is hard to believe that more recent data does not exist.

Data presentation and analysis shown in figure 5.9-21 would be much more useful if it involved measurements in individuals rather than population means. It should be an ANCOVA with size as a covariate to assess if [Hg] differs among sites and years given the underlying relationship with body size. If set up as a repeated measures the time trend could be tested. As it stands the plot infers that the reason for increased [Hg] is fish body but the interesting question is how it varies among sites and over time.

Jackson Lake was also sampled for fish [Hg] and data contrasted to human consumption guidelines. As above, analyses should be on individuals in figure 5.12-20. The analysis shows stronger relationships with age than length so should be shown. The regional [Hg] data shows high variation, likely due to a combination of variation in local exposure and variation in trophic ontogeny through size and age classes.

Naphthenic Acids seem to be prevalent in the basin. Are there any ideas of potential effects on biota and would it make sense to develop targeted studies to assess if these effects are observed?

There appears to be evidence that arsenic in water is higher in test than baseline in Cluster 2 and 3 systems. What are the impacts of arsenic in biota in general and are these observed or explored?

Total Hydrocarbons and PAHs tend to be higher at test sites. What are the impacts of arsenic in biota in general and are these observed or explored?

The biggest weakness in the fish tissue analysis program is its spotty spatial coverage and exclusive focus on large-bodied species of interest for human consumption (important for those consuming fish locally but not useful in pinpointing contamination hotspots). The program could be much more effective in addressing the RAMP objectives by sampling throughout the basin and including small-bodied, less mobile species that better represent local toxicological conditions.

Community Metrics

A pilot project was conducted to assess the feasibility of developing an integrated index of the magnitude of oil sands development effects on fishes in the Athabasca River basin. I think that this is an excellent idea and could provide an integrative index for long term assessments as the basin develops. Of course as it is currently presented the results are circular – the data is used to identify components that are best at identifying development effects and then an assessment of the resultant index shows effects, not a criticism, just a note to make it clear that nothing has been proven at this stage other than there are metrics that might be diagnostic. There are several take home messages from this pilot: (a) many more sites are needed, (b) bigger sites are needed so more individuals are captured, (c) the program must identify maximum impact and minimum impact sites so the index is appropriately scaled along this

gradient, (d) integrate the chemical, physical and hydrological and benthos components similarly and create a metric that uses the most discriminating of each of these components. This last comment could be addressed using canonical correlation to identify the most important axes for each of these components.

Field sampling for the development of this index should be stratified by habitat type and have at least 30 sites per strata. The best index would include sites outside the oil sands area to incorporate the maximum range of community metrics

Non-fish RAMP Components

Sediment Toxicity – the statement "no consistent differences in survival are apparent with respect to location" is not consistent with the figures. Show ellipses for test and baseline sites. Why not show directly the relationship between sediment chemistry and survival (a correlation/regression type of plot with measures of contaminant concentration or serial dilutions).

Sediment-benthos relationships – should use correspondence analysis to ask if you can explain benthic community composition with a suite of sediment environmental and toxicological gradients. Authors conclude "'depositional nature exerts a stronger influence...than concentrations of hydrocarbons, PAHs or metals". Of course this is the case in all but the most contaminated situations. The nature of environmental limits and controls on the distribution of biota must be strong or everything would exist everywhere. The key question for RAMP is "do the contaminants reduce the success of biota in the environments where they exist?"

Benthic Community – this work provides a seemingly robust analysis of test versus baseline conditions using a regional baseline mean and variance to assess the "significance' of deviations at test sites. The concern that appears to be recognized in the report is that the regional baseline confidence interval is in part due to variation not stratified for. The erosional-depositional stratification is real and important but are there other habitat characteristics that

are also important? If so the baseline confidence interval is inappropriately broad leading to a bias to conclude no effect when in fact there is an effect. A useful example of this is the sculpin sentinel species project in which stratification by temperature is absolutely necessary to understand the patterns. As I have discussed elsewhere, this argues strongly for the development of mechanistic models of hydrologic, chemical and biotic constraints on benthos abundance by species (or guilds) which then becomes the baseline against which test sites are contrasted.

General Recommendations

- Organize the Technical Report as a cumulative living document in which data and analyses grow with each subsequent year. It is difficult to continually have to refer to earlier documents to get the whole picture on a particular topic. In addition, organize the analysis by topic rather than by river/site. The fish section is dispersed over many different spatial locations and in a Regional Synthesis and this inhibits identification of general patterns. The sentinel species analysis, hidden within the Steepbank River section, is a good example of how a thematic rather than spatial presentation is more effective.
- Conduct age-structured demographic analyses wherever possible to estimate rates of growth, survival, recruitment. If there are industrial development impacts on fish populations they will be manifest in changes to one or more of these demographic processes.
- Continue the Muskeg River fish fence spawning survey in all years with sufficiently low spring discharge. Also extend the spawning fish fence program to other trap-able tributaries. This provides key information on reproductive effort, recruitment, growth and survival for the Athabasca Basin over time as it is developed meeting one of the key Objectives.
- There are three serious weaknesses in RAMP that must be addressed if the 3 primary objectives are to be met. I have raised them in relation to various components above, and will now generalize:

1. The first is the general approach of the control-impact design. This works well if the inferred processes are clear, and endpoints clearly directional. This is not the case here. Ecological systems like this are complex, the processes of impact are not well understood and the spatial and temporal scales of impacts and zones of influence of the various biota are poorly understood. In this case a simple contrast of a measure above and below a putative impact is black-box (i.e. we have no idea of what is causing any differences that may exist, or in fact if there are compensatory processes within the system that obscure impacts) leading to a low probability of insights of the real impact and any potential remediation approach. This design does permit some seemingly rigorous statistics, but little real biological insight is likely to emerge. A more informative approach involves development of mechanistic models of physical, hydrological and biological processes that control success of various species followed by application to putative impact sites to examine deviations in success. This requires several philosophical changes. First develop these models within the Athabasca basin at un-impacted sites (numbers and distributions of these will be discussed below in point 2). Second, recognize that the various components of the ecosystem are linked, in some cases strongly and in some weakly, and coordinate sampling of all components including hydrology, chemistry, and biota, both spatially and temporally. It looks as if there have been some attempts to do this in the last couple of years but this integration among components must be completed. A good example comes from the sculpin sentinel species work in which the erosional/depositional and temperature contrasts are necessary before any impacts can be assessed. Additional data on benthic prey abundance, assessments of reproductive effort and success of rearing juveniles in pristine sites would provide the models to assess impacts of development on success. Of course this is best done in the context of the fish community analysis (which I discuss below) and the sentinel species program should be imbedded within it.
2. The second philosophical shift that I will argue for is one that is front and centre in the Whittier and Hughes review. This works needs to be done at a large number of clearly stratified and random sites, not at a small number of fixed sites. If the goal of the

program was to do a quick-and-dirty assessment of point source impacts, then pairs of control-impact black-box style sampling are appropriate. My reading of the objectives is that this is to be a long-term monitoring program to follow the industrial development of the Athabasca basin over the coming decades. In this case the sampling philosophy argued for by Whittier and Hughes is much more appropriate. In this case black-box is not appropriate; we need to know how these systems work, how aspects of oil sands development perturb them, and what we need to do to remediate the impacts. This is the only way we will be able to measure cumulative impacts of industrial impact over space and time.

3. On the surface, the control=impact design seems reasonable...but only if it is clear that the source of impact is point source and unidirectional. If some unknown proportion of the potential contaminants leading to altered ecosystem function are initially air borne or in the ground water before entering the surface water then the source is not unidirectional and the upstream-downstream design is seriously flawed. If this is the case the "baseline" control sites are contaminated leading to false acceptance of the no-effect hypothesis and is therefore not able to meet RAMP objectives. I recommend strongly that a whole watershed design with random (or at least regular) sampling along all waterways from low order streams to the mainstem Athabasca River be implemented for the hydrology, chemistry, benthos and fish components in an integrated design. A spatial data base such as this could indentify "hot spots" of concern in various measures, provide time series of whole basin measures and facilitate assessments of spatial and temporal cumulative effects. The RAMP Technical Report uses the term "cumulative effects" but it is not clear how its program design can assess cumulative effects.

- The fish assemblage pilot suggests that useful aggregative metrics can be developed for long term assessments of basin wide impacts. Further work (i.e. at many sites that are stratified by habitat and random) should be explored to develop a family of sensitive indices. Embedded in these should be more detailed process oriented sentinel species approaches. I don't understand the concern with lethal sampling of small subsamples of

populations and clearly the reproductive assessments are very useful in determining species success.

- A general comment is that there is no reference in the Technical Report to the literature that is accumulating on impacts of oil sands chemicals on biota. It could be that this work would provide some very useful information that could help refine field sampling to address the stated objectives better. Also, they might be useful to design field assays that could be adopted by the sentinel species program to better assess success of species in control and impact sites and to better focus on sensitive species for the fish and benthos assemblage index development. A quick Google search raised 10 directly applicable peer-reviewed papers published in the last several years. If this literature has not been summarized for the RAMP team then I suggest that it be reviewed so that the RAMP team can be kept at the leading edge of the field.
- It seems surprising that RAMP has not developed a coordinated program assessing the impacts of environmental contaminants on critical life stages of organisms commonly used in physiological and toxicological assays. If the goal is to determining cumulative impacts, then we need to know where to look for them, and lab and field based experimental systems are a good start. Environment Canada has developed portable systems for conducting these assessments in situ with benthic organisms and fish.

Meeting Key RAMP Objectives

1. *Monitor to detect and assess cumulative effects and regional trends.*

The program does not effectively assess cumulative effects of oil sands development in the Athabasca River watershed. It must be demonstrated that the control-impact, before-design is capable of identifying changes in endpoints. It is likely that the current monitoring program is biased towards concluding no effect, even if one is present. Moving to a useful assessment of the spatial and temporal cumulative effects will require a philosophical change is design and substantially more sampling effort. A whole watershed spatial approach is much more robust in

identifying and assessing the magnitude of potential cumulative effects of oilsands development.

2. Collect baseline data to characterize variability in the oil sands area.

The RAMP program has provided key data on which to develop a rigorous monitoring program but now needs to focus on stratified random sampling to appropriately characterize spatial and temporal variability in the Athabasca watershed. If the impact of oilsands development is not strictly point source and unidirectional, then the RAMP design is not capable of measuring natural variability in the Athabasca River watershed. In fact, the focus needs to shift from the idea of variability in data to variability in processes. Objective 1 can only be effectively addressed if this change in approach is adopted.

3. Collect and compare data against which predictions contained in the EIAs can be assessed.

The EIA predictions are extremely general with no rationale for direction or magnitude of impacts so really does not offer much direction to the design of a monitoring program, industrial process or remediation.

APPENDIX H

Dr. William Franzin Review

AN ASSESSMENT OF THE REGIONAL AQUATIC MONITORING PROGRAM (RAMP) WITH AN EMPHASIS ON THE FISH COMPONENTS

Prepared for

Alberta Innovates Technology Futures
3608 - 33 Street NW
Calgary, Alberta
T2L 2A6

Prepared by

William G. Franzin
Laughing Water Arts & Science, Inc.
1006 Kilkenny Drive
Winnipeg, MB
R3T 5A5

Table of Contents

1. INTRODUCTION

RAMP has done a prodigious amount of work both this year and over the years and it is a credit to the people who make it happen that these reports are provided annually and that every few years a review such as this occurs. The program has a huge data base that needs to be made available to researchers to glean the gems of knowledge out of it. These researchers could be academics and their students or members of the Technical Committee or members of the consulting teams who produce the material. Whoever does it; this material should be sifted, analyzed and published for the good of the program, the province of Alberta and the country. The Athabasca River deserves nothing less and the data merit it.

2. REVIEW APPROACH

I reviewed this document section by section starting with the Executive Summary and working through all of the various sections picking out fish related parts. I have pointed out strengths of each area as well as suggestions for improvements in each of the Sections of the report areas as I worked through the text.

Executive Summary comments:

A comment about RAMP and the Executive Summary caveats on "focal projects". The comments on page xliii of the Executive Summary indicate a prevalent rather parochial attitude to monitoring of the regional aquatic resources of the lower Athabasca River (LAR) watershed (i.e. below the Grand Rapids). Although it is understandable that the industry members of RAMP who pay for the program apparently prefer to restrict monitoring to the areas of the river system wherein there may be direct or indirect impacts of their operations either presently or in the future, RAMP cannot reasonably be considered a regional aquatic effects monitoring program until the program embraces the whole LAR watershed. I guess it behooves existing industry members and the regulators to "strongly encourage" non-member industries in the RAMP area to "anti-up" and join what is an important regional monitoring program. Reference and "test" sites have to be selected on the basis of the best places to collect the data. Within the main stem of the LAR, that means sampling must include the whole river from just below the rapids above Fort McMurray (perhaps the only true reference area) all way downstream to include the major delta channels and all tributaries. In reality the real receiving water could be Lake Athabasca. In this report it is clear that the same old set of river reaches have been sampled the same old way as in the past 20 years with the exception of some additional sampling on the Clearwater River. The test of assemblage monitoring in the smaller rivers signals the direction for change in biota sampling in tributaries and the LAR main stem. The data collected since 1986 are valuable background data to aid in development of a full river fish monitoring program which should be a probabilistic sampling design; probably a stratified random sampling design that would provide fish presence/absence, CPUE data, and individual fish and population indicator data over the full length of the LAR. Only then will we know the distribution of species and their relative abundances along the length of the river, determine the beginnings and ends of migrations, the distribution of habitats, changes in important riparian zones and the relationship of biota to the river system. Repeated sampling will determine how many samples are required to control the variance in the data, how frequent sampling needs to take place and what

areas of the river are ecologically most important. Anecdotal information will be transformed into scientific fact. This is not an easy task but neither is it impossible; such programs have been and are carried out on even bigger rivers every year. Consider that there is about 300km of river channel; 30 - 10km blocks would each provide for 10 random 1km sample sites of which perhaps 3 per block might be sampled in any one sampling survey. Perusal of the river channel upstream of the Ft. McMurray Water Treatment Plant (WTP) on Google Earth suggests that there are perhaps 10 kilometres of river between the rapids and the WTP available as a reference reach. Over several spring, summer and fall surveys the whole river would be sampled at least once with many sites being sampled repeatedly. Sampling might include boat electrofishing, bottom trawling and beach seining as well as other methods plus associated water quality, habitat and benthos sampling.

Hydrology; page xlv: what is the cause of the Muskeg River hydrology change that sees mean winter discharge increase by 31.6% and the minimum open water daily discharge by 17.3%? I have always understood that the waters on projects were captured so it is puzzling to see an increase in discharge during periods of normally low discharge. This must surely affect fish populations. Or was 2009 just an abnormally wet year in which case this reported increase might just be an aberration?

Fish populations; considering that all but one of the KIR species are large long lived fishes it seems unlikely that they will provide any kind of early warning sufficient to allow for mitigating changes in any projects. Rather the program might do well to concentrate more on shorter-lived, smaller species (the Trout-perch is a good example presently being used) such as Lake Chub, Spottail Shiner and Flathead Chub to provide early signals of project impacts in population data. However that work will need to include detailed studies of the life history of these species in the Athabasca River. There is a paucity of data on life history of most small fish species in most of Canada. As mentioned previously the sampling needs to be much more extensive given the mobility of the species being sampled.

The table at the end of the Exec Summary (Summary assessment of RAMP 2009 monitoring results) should be accompanied by simple explanations of the meaning of each of the results where test is significantly different from baseline.

3. SECTION REVIEWS

RAMP Fish Component

General:

After reviewing the details of the Fish Component in the Technical Design and Rationale Document of 2009, I was struck by the lack of any physiological response indicators (e.g. MFO, EROD), or other potential indicators of toxicology or stress from any of the fish samples that are collected. Rather only traditional population response variables are used, with the most sophisticated being those from the EC EEM prescriptions for metal mining and pulp and paper effluent monitoring used in the sentinel species component (Table 3.33). At the same time, a whole suite of chemical measurements are made on tissue samples mainly to identify potentially unsafe levels of metals for human consumption with some assessment of them as possible toxicological threat to the fish. Those same types of samples collected in the field and put on dry ice could be used for a

multitude of analyses. I note that George Dixon and a number of EC researchers have done a considerable amount of work in the Oil Sands area on fish EROD/MFO with encouraging results.

Is the lack of use of physiological indicators deliberate due to these ongoing research activities or have results from these studies been tried and reported in the early years of RAMP or AOSERP or is it merely a result of the lack of familiarity/comfort with using these kinds of alternative response indicators on the part of the consultants who have done the work over the years? Or perhaps it is just cost/convenience. It seems unusual that the research has not culminated in some practical indicators for use by RAMP. Stress physiology was particularly suggested as an option for effects monitoring in the CEMA – IFNTTG Monitoring Workshop in March 2007. It would seem to me that use of such indicators would provide earlier warning of potential effects of oil sands discharges/activities on fish populations. They also would more clearly define reference and test sites whether they are naturally affected by tar sands or by industrial development.

I understand the Slimy Sculpin based sentinel species program may have been linked to some of the early work on enzyme inductions but it puzzles me why physiological indicators are not included in any of the present fish inventory and monitoring programs given that they may provide indication of change before any of the "population" responses such as growth and externally measurable reproductive changes (change in year class strength etc) provide any results. Genetic data to provide assessment of the presence/absence of local versus migratory stocks seems essential to all the purposes being served by fish population sampling.

I expected in this "5 Y report" a synthesis of the work to date in the Exec Summary that would explain all of the observed "different from baseline" observations, what they mean to the biology of the rivers where observed, whether or not these results are fulfilling the expectations of monitoring and if any mitigations are needed. I know some of this is in Section 8, but many people will never read beyond the Exec Summary. This report seems to be the same as those of other years rather than being a benchmark for the progress of the program.

Section 3.4 comments:

3.4.2.1: It is good to see that the fish fence on the Muskeg River is recommended to be abandoned. Fish fences are tenuous undertakings in any year and the fact that this one could be operated successfully once every three or so years during low water years made it of little value as a monitoring tool. A better way to monitor a stream that is difficult to fence is to sample with large hoop nets on alternating evenings and mornings from the beginning until the end of the run. This would allow for capture of enough fish to get the same data on length, weight etc that is collected with the fence without the added problem of maintaining the fence when conditions deteriorate. The hoop net can be deployed only when personnel are on site so it will not be at risk from logs and debris during high flows. Sampling infrequently only during low flow years probably seriously biases the data and reduces the number of species that may be found to use the river but not every year or not in low water years. It is apparent from the fence data that variation in run size and timing is quite large. One has to ask what is achieved by such a program when most of the data collected could be achieved with less investment. Unless a large tagging effort is undertaken to determine if repeated spawning is taking place (and might be reduced if the river is degraded) there is little benefit to tagging just a few fish in a run. The recommendation to cease this operation seems justified since adult fish might continue to spawn in the river even if it were quite degraded. Without follow up on

egg survival, larval drift and juvenile recruitment little is being learned that could not be gained by less intensive sampling.

It is noteworthy that 5560 white suckers were captured at the fence in 2009 however the data say that 3069 were counted moving up and 2491 were counted moving down. As indicated undoubtedly some of the downstream fish were the same as some counted going up; not a big run. There is some value in monitoring runs into streams that drain or pass through the main oil sands projects but the monitoring tool has to be one that can be used every year to get sufficient data to discover trends in the data collected.

The decline in Grayling and Mountain Whitefish in the Muskeg River over time may be a result of being fished out by anglers as access to the river improved and the population of Fort McMurray grew, or by handling or tagging damage while fish were passed through the traps over the years.

3.4.2.2 and 5.1.5.1: Athabasca Fish Inventories

The systematic sampling every year of several reaches of the LAR in the region of the oil sands operations for over 25 years continues. What is being learned by this long term sampling program in an open river with highly migratory species? I do believe that it mainly proves that these mostly large bodied fish in the KIR list continue to occupy the river, grow and breed. Probably even if the sample reaches were toxic to fish they would still be captured there while passing through. I believe this program should be rolled into a much more extensive probabilistic sampling design that would sample the whole river from below the rapids just above Fort McMurray to the major distributary channels of the Athabasca delta. Such a program could include many of the existing sample sites as well as many more in different parts of the river. Statistically, the results would be much more powerful. Another aspect that would greatly improve the program is to collect samples for DNA analysis to detect possible presence of sub-populations of the KIR species. CEMA IFNTTG did this for walleye and determined that the stocks in the river are closely related with those in western Lake Athabasca as well as well upstream of the rapids above Fort McMurray. Also it has been noted by RAMP that tagged walleye from the Fort McMurray area have been recaptured as far upstream as Lesser Slave Lake and over 100km upstream in the Peace River. It is quite likely that the other large species are as mobile as the walleye. Sampling should perhaps concentrate more on the smaller species in the river such as Flathead Chub, Lake Chub, and Spottail Shiner etc as has been done for Trout-perch. These species at least are less likely to engage in long distance migrations and may have local stocks along the river. It should be noted however, that some Flathead Chub tagged by CEMA in the vicinity of Fort McMurray were recovered in the delta so these larger cyprinids also may be making migrations in the mainstem. Finally in order to improve catchability of some smaller species, other gears in addition to boat electrofishing should be used such as bottom trawls and beach seines. These can be deployed from the same boat.

Some species of fish seem to be tagged with Floy tags at some times. In order to obtain knowledge on age and growth of individual fish the program should consider mass marking of all caught and released fish with less damaging and less losable tags such as PIT tags. These can be applied quickly to many more fish and recaptures will be obtained without many of the side effects of external tags. They also have individual numbers and can be inserted in different body areas in different years to quickly aid in identifying year of tagging at recapture.

In the summaries of fish captures by the inventory surveys it seems all the data are reported as if they only come from two locations, Athabasca River and Clearwater River when in fact there are seven separate sampling areas in the Athabasca and three in the Clearwater (this is shown in Table 3.4-2. I didn't see any results or discussion of variance in captures among the sites, some of which supposedly are near to being baseline and others test (apparently no sites in the Athabasca are considered baseline or reference). To see if there are differences among years you might have taken average catches by species by season for all years and tested 2009 against those averages to see if there were differences e.g. box and whisker plots with an average drawn across the graph. I think this may have been done in other years. Similarly species trend lines over all sampling years may have been an interesting plot. Figures 5.1-27 and 5.1-31 to 5.1-36 should all be bar graphs or scatter plots because the data on the x axes are category variables not continuous variables (this was noted in some other areas as well). In all of these bar graphs you could have computed means and done tests of the data against means to see if any years were significantly different. Another approach with the inventory sampling is to do some species accumulation curves to determine if the number of samples is sufficient to develop an asymptote of species numbers. This is a standard procedure to determine if sampling is sufficient. Another way to do that is to sub-sample the whole data base using bootstrap techniques to determine for each species how many samples are required to reach an asymptote in the numbers caught. The species requiring the greatest number of samples to reach an asymptote drives the sampling program. I can't see what information is gained by doing Correspondence analysis on the co-occurrence of species in the data tables when you can readily see which species are most frequently caught in each season and year. Statistics are hardly required for this unless you can say something significant and meaningful about the result. What does it mean to the species or the RAMP program if a species condition factor falls outside the 5th to 95th percentile? Can a proximate cause even be speculated? It is totally uncertain where any particular fish may have originated.

Obviously the fish inventory work has to be approached differently than the test/baseline approach that is used for the sentinel species component. You have an open river system with areas affected to some degree by human impacts. Usually this kind of situation would be addressed by a reference – condition approach and there is a reference area available immediately above Fort McMurray but below the rapids and some of the areas more than 100 km downstream of the projects also may approach reference condition. The test is to do the collections to find out if reference and degraded areas can be identified in the river at all, near of far from the projects. It is entirely possible that natural variation in fish metrics will overwhelm signals from industrial water use; that is the challenge for RAMP to discover.

3.4.2.3 Fish tissues:

I believe far too many metals are being analyzed far too frequently in the RAMP program. Metal assays in fish tissues are expensive and cutting back this part of the program could free up significant resources for other components. It is unlikely that metals in tissues of adult fish will change suddenly in one year so a three or five year rotation for tissue samples for metals would be more appropriate, including mercury. Also unless someone is really interested in individual fish tissue metal levels for some research project it seems unnecessary to do more than screening with composite samples of five fish from five length frequency categories for a much smaller suite of

metals than are analyzed presently. If some metal appears elevated in a composite then a more thorough sample might be analyzed. Composite samples do not require that all the tissues be used; separate individual samples can be maintained in the freezer at minor cost. Metals that are reported as undetectable are unlikely to suddenly be present in detectable quantities in fish tissue or trip any criteria of HC or EPA. In fact even if many of the metals are detectable but have no fish or human health criteria, why analyze for them annually. If they have been analyzed over a number of years with little or no change in the result clearly it isn't necessary to continue with so many analyses. Finally when you catch a fish from an open river and discover it has e.g. mercury over a HC consumption criterion, what of it? You cannot tell where that fish has been living or where it picked up the mercury. It is a data point that indicates that mercury is available to fish somewhere in the system. Plot the metals in fish tissues over the years and any that have a flat line should be deleted from the list of metals assayed. Tainting compounds of course are a bit trickier because short term changes in project activities presumably could increase potential tainting in as little as a day or two so it would be more difficult to justify changing sampling to once in three to five years for these compounds. However the monitoring program is not designed to capture events such as spills so perhaps even the tainting analyses are more numerous and frequent than they need to be.

Unless the regional lakes are subject to potential effects of air pollution by being in the airshed of the Oil Sands probably RAMP should leave the sampling of regional lakes to the province and/or Health Canada. There are enough ways to use the consultants' time and the industry's funds without doing what should be a federally or provincially funded program.

5.3.5.2 Sentinel species:

The Athabasca River Trout-perch Sentinel Species program appears to be fully integrated into the Athabasca River Fish Populations (KIR) component. I searched the whole pdf of the 2009 report and didn't see any data on the EEM types of measurement endpoints for Trout-perch. Are these data still been produced from samples collected in 2009? Also it is possible that developing some suitable physiological indicators such as MFO/EROD assays might detect potential changes in Trout-perch biology ahead of the more physical measures including fecundity and egg size. It would be good if it were possible to develop an index of YOY year class strength perhaps at 1+ age (these will have passed the test of first overwintering). Alternatively analyses declining year class strength over time would provide an early indication of population level effects. This however requires some tests of equal catchability by size/age. Also some demonstration of the assumption that Trout-perch are to some degree sedentary in sections of the river is needed. In spite of their ubiquity in Canada, Trout-perch are relatively poorly studied, like most other small bodied fishes. However they are known to have a protracted spawning period and like Spottail Shiners, fish in many lake populations migrate into streams to spawn.

I suggest that the Slimy Sculpin sampling program is not sampling YOY fish at all unless there are age data to back up the claim of YOY fish reaching 50mm in October. I think it is unlikely that 50mm fish in October are YOY. I checked for some Slimy Sculpin ageing data via the internet and found a study in northern BC (Carmichael and Chapman, data courtesy Bruce Carmichael, pers. Com. BC Environment) which included ageing work with otoliths. In Martin Creek, BC (just north of Dawson Creek), 2+ fish in October were 46-59mm in length and a 3+ fish were 58-79mm in length (no 1+ fish lengths were reported). Sculpins taken in the lethal sampling or incidentally to

other programs should be aged by otoliths to develop an age length key for the species. My experience with capturing several sculpin species in small streams is that YOY in late summer are in the range of 25-30mm in length and would be very unlikely to reach 50mm until well into the 1+ year class. YOY sculpins are very difficult to catch, or even see, when shocked since they just roll over wherever and stop moving. In a rough stream bottom they will be deep in the crevices between stones. In the account of sculpin collection on page 5-175 I believe relative abundance is confused with CPUE but then effort is not given. It is assumed that effort is standardized across sites and sampling times (suggested but not stated in the TDR report) but it is not stated. One could collect estimates of population size by barrier netting the sample site, doing triple pass electrofishing and using removal method statistics (e.g. Zippin method). Although non-lethal sampling is used in the sentinel species program it should be noted that electrofishing sculpins may cause significant spinal damage especially with multiple sampling in the same place (see Clément and Cunjak. 2010. NAJFM 30:840). Another point about the sentinel program is that it seems that two baseline sites is unlikely to be sufficient for such a program, a minimum of three and better about 5 should be sought. This is especially true in this program where the differences between baseline sites are greater than the differences between the baseline and test sites.

Section 5.9.5 Fish populations:

Note an error on page 5-323 2nd last paragraph; it says species richness was greatest in the fall when Figure 5.9-8 and Table 5.9-15 both indicate it was in summer. It is worth noting that the capture efficiency in the Clearwater was rather poor with almost as many fish seen and not caught as were caught, especially for Northern Pike and Spottail Shiners. This places CPUE in some doubt and causes concern about the gear used and the capabilities of the netters. How is electrofishing standardized between years? Is annual training in electrofishing and dipnetting part of the QA/QC for the fish inventory program? Comparing between years is confounded by fish movements, timing, hydrology and the gear and netters each year. On page 5-325 re: length frequency of Northern Pike one must wonder if mainly large ones or small ones escaped (were seen not captured) and what influence that observation has for the analysis. Once again relative abundance is confused with CPUE. The length frequency distributions that we see in catches at one point in time will never be that close to that of the whole population but it can be approximated with sufficient sampling. Page 5-327 Summary; these samples mainly represent presence absence with some understanding of relative abundance among the species present during the fishing period. Clearly we need a better understanding of the "home ranges" of many of these species and life stages. This will be achieved only by much more extensive sampling, intensive tagging with perhaps PIT tags and more DNA work to help understand stocks involved.

Section 6.2 Fish assemblage monitoring pilot study

It is nice to see a more scientific approach being tested in the RAMP program. I realize this is a significant deviation from past practice but it is a welcome change. Page 6-42; discussion; we should be glad that at this point none of the test reaches in the rivers examined have become seriously degraded. If they can be kept so, the whole Oil Sands industry should be gratified and satisfied that the money spent on RAMP and CEMA is justified. Page 6-44 point 4; there still is time to redesign the program and redirect the RAMP program funds from other activities to improving on this pilot study. One significant outcome could be a set of northern Alberta boreal forest fish IBI metrics that will usable across the northern part of the province and perhaps into BC, SK and NWT. Bravo for

finally taking the plunge into a scientific redesign of the fish component and the integration with other components.

Section 7.4 Fish populations;

There is a great urgency to data mine the existing RAMP database to learn what has been found in much more detail, species by species. Ageing needs to be completed for all of the samples in a timely manner. There is opportunity for desktop research for a multitude of graduate and undergraduate theses in this database with potential great benefit to RAMP. Page 7-46; probably the analyses of mercury in fish tissues should be passed to HC with RAMP just providing the samples on something like a five year cycle. It is HC's responsibility to monitor the health of country foods in Canada, not RAMP's. As indicated in point 4 the lakes are mostly not in the RAMP area anyway.

Section 8.4 Fish populations:

Page 8-7 Fish inventory; the Summer inventory on the Clearwater should not continue as in 2008-09 but rather be rolled into a broader Athabasca River assemblage monitoring program as was done on the smaller rivers in 2009. It is time for a change.

Page 8-8 Sentinel species; some consideration should be given to improving this component so that there is at least 4 or 5 baseline sites in the mix. Also ageing needs to be done to be sure what exposures fish have and what the lengths presently reported represent in age and time of exposure.

Section 8.4.2:

1) The inventory should be revised to be more extensive and more statistically acceptable as a monitoring program.

2) Pathology should be included in the assemblage monitoring program along with habitat, benthic organisms, etc as in the pilot study.

3) The addition of physiological indicators (check with George Dixon and others on this) to the sentinel species program should be considered since they are likely to provide indications of change or presence of deleterious conditions prior to population characteristics like L/W, condition etc.

4. Final Comments

I read briefly through the fish section of the last review report by Post, Munkittrick, Dubé and Souter as well as the report from Whittier and Hughes and it is a bit disturbing that other than the recommendation from Hughes and Whittier to try assemblage monitoring and the use of Slimy Sculpin and Trout-perch instead of large bodied species for sentinel species there have not been many changes in the five years since that last review report. Many of the comments provided in that report still apply. One can hope that change will start to occur more rapidly. Examples of easy changes were the recommendations that RAMP engage a Scientific Advisory Board and that the database be made publicly available. I believe that RAMP could benefit from input from recognized scientific experts in monitoring programs on an annual basis. The members of the RAMP Technical

Committee and the Steering Committee all have busy full time jobs in their parent organizations (I know this from personal experience) and cannot be expected to be experts in biomonitoring as well. These people provide a vital link to the real world of the Oil Sands area and the LAR but should not be expected to design the best possible monitoring program for such a large river system. A ten year synthesis of what has been done and what has been learned by the monitoring that has gone on since 1997 (not another annual report comparing past years with trend data but a true synthesis) would be a first step in a process for change.

A final plea: Obviously preparing the 2009 RAMP report is a really significant undertaking and it is a laudable achievement. However, the present structure of the report is a reviewer's nightmare as long as the reviews are going to be done by faunal/ecological area experts. It took me more than two hours just to go through the entire report to locate all of the fish related sections in the report and yet more time to find all of the associated tables and figures and then to collate it into a collection of sections for review. It might be better in future reviews if the report was broken into the disciplinary sections so that all material (methods, results, discussion) in one subject area would be together in a chapter with an integration chapter after the various component chapters. If the whole program moves toward assemblage monitoring this will be easier because they will be more closely integrated both in the field and in the report. It seems unlikely that many people will read through the entire report in a comprehensive way due to the complexity of the program; it is more likely that disciplinary experts would focus on their own areas in such a report. I do realize there are some advantages to the present layout for the writers but just going to the Fish Component in the document would miss a great deal of other fish related work. Perhaps for the review a different format could be used with cross referencing where needed.

APPENDIX H1

Addendum to Appendix H

Addendum to Appendix H
An Assessment Of The Regional Aquatic Monitoring Program (Ramp) With An Emphasis On The Fish Components

Having had the chance to read some of the comments of the other reviewers before putting down my own thoughts takes away some independence in thinking but then there wasn't much of that before now anyway. I have to agree with much of what Kelly and John have written as well with Cathy's notes from her conversation with Joe. It gave me cause to reflect on what RAMP really is. I think the history of the program is largely responsible for what its outputs have been and thus responsible for the comments that each of us has had. This is my understanding of the history. If you recall some comments that Terry Van Meer made at the beginning of the first meeting we had and some he made at his "RAMP retirement", you will understand my thread. He said that he (from Syncrude) and a biologist from Suncor (can't remember his name now) in the early days of oil sands development recognized that they were doing a lot of duplicate sampling to answer requirements in their operating licenses for monitoring of predicted impacts of their operations on the environment. In the aquatic realm early on that was mostly about the main stem of the Athabasca River and the streams that ran off of the original Syncrude and Suncor leases. A combined program was initiated under what became the RAMP banner and each time a new oil sands operator began development of a new lease they were encouraged to join RAMP by industry and government alike to provide a centralized "monitoring agency". So RAMP grew to its present form incrementally somewhat like a rapidly growing city without a planning department and no city plan. I am sure all of you have seen the result of poorly planned urban sprawl; greater Phoenix comes to mind. That is the RAMP monitoring program; a history of many competing demands from too few resources (financial and human) have resulted in the sort of wide ranging program we see now. Having said that, for the last two reviews, reviewers have struggled with what exactly RAMP is supposed to be monitoring and have made recommendations to improve it. Recommendations made have not always been adopted (probably mainly for lack of committed resources and inertia) and even those that have been adopted have been adapted to the program in place and usually are test run before implantation e.g. the Clearwater River sampling reported this year. Implementation of change in RAMP takes a long time because of the lengthy budget request and allocation process run by an industry committee that really would rather not spend money on monitoring. A recommendation accepted in 2010 probably would not see implementation until 2012 at the earliest. I worked with CEMA for 10 years and the budgeting process and lag times were the same.

It has been mentioned a number of times, both in past reviews and during the present review that governments (AENV, DFO, EC) should be involved in monitoring along with industry. Of course government sits in the committees of RAMP and CEMA but they seldom commit any resources to the programs. I think if you look at any of the big monitoring programs in the US or even in the St. Lawrence River Program in Canada you will find a major government involvement (due to international, interprovincial or interstate concerns) (e.g. EPA, USGS, EC) as well as state/provincial governments. The same kind of problem exists for Great Lakes; the Laurentian great lakes get plenty of attention for the same reason and for the same reason lakes like Winnipeg, Great Slave and Great Bear get virtually no attention. Therefore part of RAMPs's problem is that this whole development is within one province (also the main beneficiary of the resource but

Canada gets plenty). The fact that downstream jurisdictions (First Nations and the NWT) are affected by what goes on in the oil sands area should invoke federal government mandates to participate fully in a monitoring program but it has not. Without interstate, interprovincial or international issues, those other large monitoring programs would be like RAMP, an evolved program with no clear plan that is unlikely to be sensitive to early recognition of impacts. Therein is the basis of RAMP monitoring and the basis for the criticism that reviewers aren't sure what RAMP is. Is it surveillance, monitoring of development effects locally or regional monitoring; at present RAMP is trying to do some of all of these, addressing too many differing mandates with no specific question. We really need a strong steady hand of government to be involved in oil sands monitoring.

The solution? First integration of the other ongoing reviews is necessary to make sure all of the interests are represented. Then, I think it is time to have a broadly based stakeholder conference or workshop to decide what should be monitored by whom and how in a transparent open process. As Whittier and Hughes point out in their major recommendations the data in hand is useful for scoping exactly what should be done with a good deal of the how thrown in. Is there the collective will of industry and government to seriously address this? I don't think so but I would love to be wrong.

A first step in this process would be a real synthesis of all of the RAMP data (and earlier material from AOSERP) collected so far with a critical evaluation of what parts of the program over the last 10 years have provided data suitable for the assessment of ecological change and the potential influences, natural or anthropogenic. With that in hand an astute group of experts in monitoring might be able to design a program from the ground up that will over the next 10 years be able to detect change, causes and impacts. I assume the monitoring recommendations from CEMA for IFN in the Athabasca will be of interest to RAMP and that the companies and government will realize doing two separate monitoring programs will be a duplication of effort and therefore work to develop a single comprehensive monitoring program. There really are not enough resources all around to do both. I think a complete program could be put in place within 2-3 years if there is a will to see it happen.

APPENDIX I

Dr. S. Watmough Review

2010 RAMP REVIEW WITH EMPHASIS ON THE ACID SENSITIVE LAKES (ASL)

Shaun A. Watmough

Associate Professor, Environmental and Resource Studies, Trent University,

Peterborough, Ontario, K9J 7B8

swatmough@trentu.ca

Tel. 705 748 1011 (7876)

September 5, 2010

Table of contents

1. Introduction and Review Approach

I was contacted in early August 2010 to inquire whether I would be willing to review the Acid Sensitive Lakes Component (ASL) of the Regional Aquatics Monitoring Program (RAMP) and was provided with the information on or about August 18, 2010. Given my late involvement in the process, I was unable to attend the meeting in Vancouver where the procedure was described, but I was given meeting notes that provided a summary of the RAMP program and outlined what is expected in the review process. Consequently, this review follows the guidelines and answers the questions that are outlined below.

The reviewers were asked to evaluate as to whether the current RAMP program is meeting the following objectives (outlined in the Design and Rationale document):

1. Monitor aquatic environments in the oil sands region to detect and assess cumulative effects and regional trends.
2. Collect baseline data to characterize variability in the oil sands area.
3. Collect and compare data against which predictions contained in Environmental Impact Assessments (EIAs) can be assessed.
4. Continuously review and adjust the program to incorporate monitoring results, technological advances and community concerns, and new or changed approval conditions.
5. Conduct a periodic peer review of the program's objectives against its results, and recommend adjustments necessary for the program's success.

'Each reviewer is required to review the whole program with emphasis on recommendations to the component under their expertise. The review should include recommendations to the other components that would be required to support the component under the reviewers' expertise.'

In my case the component under review is the Acid Sensitive Lakes (ASL). In order to complete my review I read through the RAMP Technical Design and Rationale Report, the 2009 Ramp Annual Report in detail and referred to several other documents including the

RAMP Terms of Reference, all the data associated with the ASL and other RAMP reports contained within the RAMP website including the 2005 RAMP report.

2. RAMP Program Overview

The RAMP program is very challenging and overall it is a very impressive program that has been in operation since 1997. The program seeks to monitor changes in hydrology, water quality, benthic invertebrate communities, sediment quality, fish populations and acid sensitive lakes. The overall objective of the RAMP program is to answer three questions:

1. Can the present Program detect changes if they occur?
2. Is the present Program appropriate for identification of potential sources resulting in the change(s) if found present.
3. The appropriate questions being asked by the Program and the appropriate criteria being monitored to answer those questions?

To begin, I would like to comment on these three questions as they relate to the Program as a whole.

The answer to question 1, as written, must be NO. The program is conducted in a large and highly heterogeneous area with very limited 'baseline' data and for most of the components several measures are taken in order to look for change/impacts. The reality is that for many of these components there could well be change that is not detected in the monitoring Program that is simply not detectable or is not considered detrimental. For example, by definition and change in hydrology of less than 5% is not detectable and there may well be changes in other components that fall within the natural variability. This is not meant as a criticism, rather that the question should be better defined; for example 'Can the present Program detect changes '**that are considered detrimental to ecosystem health'** if they occur? – or something similar.

Once this question is defined, it leads on to my next point regarding which criteria are used in the assessment. In both reports (RAMP Technical Design and Rationale; RAMP 2009

Annual Report) a number of measures are taken for the various components and yet it does not appear that they are all used in the assessment for various reasons. I found this to be somewhat confusing and believe that it should be clarified. Looking at the Summary Assessment of RAMP 2009 Monitoring Results the following parameters were used as basis of the Assessment:

Hydrology: Calculated on differences between observed *test* and estimated *baseline* hydrographs: ± 5% - Negligible-Low; ± 15% - Moderate; > 15% - High.

Water Quality: Classification based on adaptation of CCME water quality index.

Benthic Invertebrate Communities: Classification based on statistical differences in measurement endpoints between *baseline* and *test* reaches as well as comparisons to regional *baseline* conditions.

Sediment Quality: Classification based on adaptation of CCME sediment quality index.

Fish Populations (fish tissue): Uses various USEPA and Health Canada criteria for risks to human health, fish health, and tainting from fish tissue concentrations of various substances. LKWH-lake whitefish; WALL-walleye; NRPK-northern pike
Fish Populations (sentinel species): Uses Pulp and Paper Environmental Effects Monitoring Criteria (Environment Canada 2005).

Acid-Sensitive Lakes: Classification based the frequency in each region with which values of seven measurement endpoints in 2009 were more than twice the standard deviation from their long-term mean in each lake.

Consequently, the above criteria are the ‘changes that can be detected’ as listed in the first question and which are presumed to relate to level of change that is undesirable. For some components, such as hydrology, this may well be true, but for other components (ASL which will be documented later in this report) it may well not be sufficient. This also raises the question of why so many other parameters are measured and yet do not appear to be used in the assessments. It would be much clearer if the documents stated very clearly from the

onset what measure of change is being assessed for each component and the rationale for this measure. It should also clearly explicitly state the reasons for other measures and how they will be used in the assessment.

The answer to the second question asked of the reviewers regarding identification of the sources depends on the component being asked. My expertise is on ASL, and in my opinion sources that may cause changes in water quality cannot be identified within the RAMP Program. Unlike all the other components within the RAMP Program, the ASL is not watershed based and lakes are only impacted by anthropogenic deposition. Therefore in order to attribute source to any changes in lake chemistry an atmospheric model is needed, which does not appear to be part of the RAMP Program. Furthermore, any lake will receive anthropogenic deposition from multiple sources (including regional and long-range transport) and so there will almost always be multiple sources. Clearly, the approach used for ASL is very different from the other components and I did not feel that this is clearly expressed early on in Section 1 in the Technical Design and Rationale Report. For all other components there are test and baseline (or estimated baseline) areas, which is not really the case for ASL and this should be clarified early on in any documentation.

Whether the approach used in RAMP for the other components (hydrology, water quality, benthic invertebrate communities, sediment quality and fish populations) is best answered by the specialist reviewers in these areas. My opinion is that most test site are located downstream from potential sources and so the answer is probably yes, although again there may well be some discrepancies among components (i.e. fish populations are very heterogeneous and are not restricted to an single watershed). One issue that arises, but is probably better answered by the specialist reviewer is whether location of the test sites is appropriate to detect source?

The final question relates to whether ‘appropriate questions being asked by the Program and the appropriate criteria being monitored to answer those questions’. Part of my answer to this question was given above when answering the first question. It is clear from section 2 in the RAMP Technical Design and Rationale Report that the Program is heavily influenced by results from the numerous EIAs that have been conducted in the region and in which multiple criteria have been assessed. A common approach in EIA’s is to conduct several quantitative

assessments and convey these measures in a risk index, i.e. Low, Moderate or High and this approach is used in the RAMP program. Each EIA asks questions that are generally specific to that assessment although there is a lot of overlap in certain areas. The expected effects of oil sands operations are outlined in Figures 2.1 and 2.2 of the RAMP Technical Design and Rationale Report. I suggest these figures could be improved as it is not entirely clear why the figures for surface mines and *in situ* projects are so different. Impacts from both operations are associated with activities and introduction of contaminants and although these may differ in nature, the general pathways should be the same. For example, changes in watercourse discharge and water levels has impacts on fish habitat in Figure 2.1, whereas surface water hydrology has no impact on fish habitat in figure 2.2.; there are no wetlands in Figure 2.1, but there are wetlands in figure 2.2; Contaminants impacts biodiversity directly in figure 2.2, but not in figure 2.1; surface water quality does impact BMI in Figure 2.1, but does so in Figure 2.2 etc. These figures need to be revised.

Table 2.11 lists the measurement endpoints used in Athabasca oil sands projects and table 2.12 lists the criteria for assessment. I would like to make two small points here. Firstly, the two tables are not directly comparable as different names for specific projects are used (i.e. Suncor millennium is referred to under Climate and Hydrology in Table 2.11, whereas Suncor Firebag and Opti/Nixen Long Lake are referred to in Table 2.12 etc.). It may well be that the same thing is being referred to, but the tables should be checked for consistency. In section 3 of the RAMP Technical Design and Rationale Report, there is much reference made to the results of the 17 EIA's and the fact that for many of the predictions associated with various components yield negligible or low impact. When doing this I assume the report is reporting the results from the EIAs directly, however the different EIAs often use different criteria for negligible, low, moderate and high (Table 2.12). For example, when assessing water quality - a change of 10% in the measurement endpoints in the Canadian Natural Horizon project would be considered Low, whereas a change in endpoint of 5% in the Opti/Nexen Long Lake Project would be considered moderate. As Opti/Nexen Long Lake Project is not listed in water quality in Table 2.11, I have no idea if the same endpoints are used. My suggestion is that RAMP explicitly acknowledges the fact that summation of results from EIAs may be somewhat arbitrary for certain parameters or attempts to standardize when possible.

These relatively minor points aside, I believe that the RAMP program is generally asking the appropriate questions and using the appropriate criteria although it would be easier if the main assessment criteria are explicitly stated early on in the document. However, I believe there is need for improvement/change, particularly with respect to ASL, which is outlined below.

Other minor points associated with the RAMP program and reports that I have read are:

1. Acronyms: KIR should be Key Indicator Resources; SBC is usually sum of base cations, not the ratio of alkalinity to base cations
2. Quality of Figures/Tables. While the tables and figures are very good, the legends/titles are often very poor, particularly in the RAMP Technical Design and Rationale Report. For example in Figure 1.2, I assume the flow values are annual flow based on estimated baseline or are they measured? In Table 1.2 I assume capacity is barrels per day? Etc.
3. The terminology used in the reports is often a little confusing. For example, in section 1, when describing the RAMP components on page 1-2, rivers and creeks are referred to in climate and hydrology and in benthic invertebrates rivers, streams and wetlands are referred to. In section 3.4.5.1 that describes monitoring protocols for hydrology, stream flow is referred to and in section 3.6.5 wetlands are analogous to shallow lakes. Standardization of terminology throughout would be good.
4. The calculation on page 3-79 is not clear to me. If I put the numbers given into the text into the equation I get 7.1 not 1 as box 1 suggests.

3. Acid Sensitive Lakes: Strengths

The acid sensitive lakes (ASL) component is considerably different to the other components assessed in the RAMP program for a number of reasons including the fact that it is not confined to a particular watershed, is impacted only by regional stressors and does not try to compare baseline with test conditions. Any recommendations that I make for this component are therefore independent of the other components assessed by RAMP. In addition, all the EIA studies that evaluated ASL appear to use critical loads (or exceedance of the critical load (2-22) [I am not sure how acid deposition is a measurement endpoint reported for Syncrude

Aurora on 2-22], whereas in the 2009 RAMP report a 'classification based the frequency in each region with which values of seven measurement endpoints in 2009 were more than twice the standard deviation from their long-term mean in each lake' was used. This endpoint is not used in any EIA and although critical loads and exceedances are calculated, they do not appear to be used for assessment purposes given the fact that a large percentage of lakes exceed the critical load (as calculated) but are reported to have negligible impact.

The strengths of the Program are listed as follow:

1. A relatively large number of lakes have been continuously monitored and between 47 and 50 have been monitored for 8 years, allowing trend analysis to be conducted. [Note the errors in table 3.39; the data are from 1999 not 1997 and more than 50 lakes are presented as 5 were dropped after the first year; this is not clear in the supporting text]. This number of lakes is consistent with other acid water monitoring programs for the size of the region.
2. The measurements are taken during fall turnover, which is the appropriate time to take water chemistry measurements and should reduce the natural year-to-year variability as much as possible. Ten lakes are sampled more regularly to examine seasonal variability, although this does not appear to be taken into account in any assessment (i.e. just something to be aware of).
3. The RAMP program analyses all the appropriate chemical parameters needed to assess the acid sensitivity of lakes.
4. All analyses follow consistent protocols and adhere to acceptable QA/QC procedures although in the database there appear to be some anomalous values for some lakes in some years (i.e. L107, 2007; Ca is almost double the value of all other years; same comment for E52 in 2005). Is there a possibility that data from some lakes have been mixed? Also, it is not clear why Ca-T is sometimes quite different from Ca (*i.e*. L107, 2005).

4. Acid Sensitive Lakes: Potential Areas for Improvement

I believe that the RAMP ASL Program could be improved in a number of ways that are documented in detail in the Appendix. The majority of my suggestions relate to interpretation

of the data and I feel are readily accomplished, but I do have a few minor comments regarding sampling design. The sampling design (choice of study lakes) depends entirely on what question is being asked and how the data/results will be used. It is clear that criteria for lake selection (3-147) have changed over the years. Selection has focused on ‘acid sensitive’ lakes although the criteria of having a total alkalinity of less than 400 µeq/L is double the value used by Environment Canada, for example (Jeffries et al. 2010), a range in DOC, potential acid deposition and representative of the physiographic sub-regions. The more recent additions essentially added lakes in areas that correspond to the region of greatest oil sands activity and with high potential critical load exceedance. This sampling design is clearly not a representative survey of the lakes and the data cannot be used to assess the potential impact of acid deposition on a regional basis (Jeffries et al. 2010) as it would likely overestimate the risk. Given the fact that acid deposition is a regional issue, this may be a concern. While the sampling design is clearly not representative of lakes in the region, it is typical of many lake monitoring Programs in that it targets the most acid sensitive lakes as these are most likely to show effects and should be considered when interpreting impacts. Therefore the rationale for the choice of lakes and how changes will be interpreted should be explicitly stated.

The lakes were chosen to assess the impact of acid deposition. In the RAMP Program, PAI (S + N – BC) is used, which is not the same as acid deposition. Also, in contrast to other components in which all potential impacts are considered, the ASL component focuses on acidity (metals have also been added) – the potential for eutrophication (excess N) is not considered. Given the fact that N emissions in the oil sands may well exceed S emissions, this may be an oversight. The chemical measures currently adopted would be able to identify potential eutrophication impacts – although choosing only acid sensitive lakes may not necessarily reflect an appropriate design for detecting eutrophication effects.

Changes in lake chemistry are based on single fall values, although seasonal sampling is conducted for 10 lakes. Some studies in Boreal regions have indicated that episodic effects associated with snowmelt in particular may be more severe than chronic long term changes (Laudon et al. 2004). Seasonal data should be evaluated to see if a) they are appropriate fro detecting episodic effects and b) if they are, is there any evidence that episodic effects occur.

The basic objective of the Program is to ‘detect effects of acidifying deposition on water quality and lake biology’. All the data I found relate only to lake chemistry – there is no biological data because this is collected by other agencies (Alberta Environment/Environment Canada). An effort should be made to link the chemistry to the biology. If it is not in the RAMP program, impacts on Biology cannot be assessed.

However, my main suggestions pertain to the interpretation of the results and the overall assessment. Within the RAMP program four primary data analyses are conducted as described in the RAMP Technical Design and Rationale Report.

1. Between-year comparisons of measurement endpoints over the entire population of lakes – Given the high natural spatial variability in lake chemistry and the fact that lakes potentially receive different levels of acid deposition and each lake will respond differently, I feel that this analysis has limited value for identifying impacts. I feel that potentially harmful changes in chemistry in some lakes will not be identified using this approach.

2. Calculation of critical loads of acidity and comparison to modeled potential acid input has limited value because of the way it is done. Firstly, conceptually it is wrong as the SSWC model is a steady state model and therefore the critical load should not change over time. It need only be calculated once (using multi-year average chemistry) and exeedance can be calculated over time as deposition changes (Henriksen et al. 2002). It is recognised that the critical load estimated by the SSWC can change over time (Rapp and Bishop, 2009: Watmough et al. 2004), but this is because the F-Factor can change as soils acidify which is not captured in a steady-state model and which is why dynamic acidification models are used (Rapp and Bishop 2009). Finally, there are a number of other issues associated with the calculation of the critical load (BC_o, contribution of organic acidity, appropriate chemical limit, runoff, negative critical loads) and exceedance (use of PAI) which are problematic and are outlined in detail in the Appendix.

3. Mann-Kendall trend analysis on measurement endpoints in individual lakes. This is by far the most common way in which acid deposition impacts on lake chemistry are assessed (see Burns et al. 2006: Davies et al. 2005: Evans et al. 2001: Skjelkvale et al. 2001: Stoddard et al. 1999 as examples). These data are presented in the annual report and in my opinion should

be the way in which impacts of acid deposition on the lakes are assessed (i.e. # lakes showing trends in chemistry that are consistent with acidification impacts). Using this approach, it should be confirmed that the lake selection is appropriate for the question being asked (i.e. lakes are not representative of each region; they are targeted for acidic effects, not eutrophication). Also it may be worth calculating ANC (SAA-SBC) for the lakes and using this as a chemical endpoint, given the fact that it is used in the critical load calculations.

4. In addition to 3, RAMP uses Shewart control plots for 10 control lakes deemed most at risk for acidification. These 10 lakes are selected based on the ratio of the PAI to critical load. However in the 2009 annual report the summary for the potential for acidification was calculated for all lakes in a way that was not described in the RAMP Technical Design and Rationale Report.

> ‘For each lake, the mean and standard deviation were calculated for each measurement endpoint over all the monitoring years. The number of lakes in 2009 within each sub-region having measurement endpoint values greater than two standard deviations (SD) (above or below the mean as indicated above) was calculated. The number of such endpoint-lake exceedances was expressed as a percentage of the total number of lake-endpoint combinations for each sub-region. The results were classified as follows:
>
> Negligible-Low: sub-region has <2% endpoint-lake combinations exceeding ± 2 SD criterion;
> Moderate: sub-region has 2% to 10 % endpoint-lake combinations exceeding ± 2 SD criterion; and
> High: sub-region has > 10% of endpoint-lake combinations exceeding ± 2 SD criterion.’

It is not clear to me why there is this discrepancy between the two reports. Furthermore, I do not agree with the use of Shewart charts for three main reasons:

1. The calculation of standard deviation appears to use all the data collected for a given lake (i.e. not restricted to ‘baseline’ (unless I missed something)) and so standard deviation will change over time as data are added. Any increasing trends in a chemical parameter will be incorporated into the calculation of standard deviation, thus limiting the ability to detect change.

2. For several parameters, calculation of 2 or 3 SD will result in negative values (i.e. Fig 7.5-2 in the 2009 annual report).

3. The amount of change associated with 2 or 3 SD for many chemical endpoints is beyond the range of observed chemical changes in many other regions. For example, in Figure 7.5-3, 2SD corresponds to a change in pH of more than 1 pH unit.

5. Recommendations for ASL

Overall, I suggest:

1. Clarifying how ASL lakes are assessed for potential acidification impacts;
2. Critical load calculations and the use of critical loads should be re-evaluated as they are potentially useful, but are not currently calculated in a scientifically defensible way.
3. Shewart Charts (or values exceeding 2 or 3 SD) should not be used as a primary means of assessment.
4. The number of lakes showing trends consistent with acidification should be the primary means of assessment for potential acidification impacts for the ASL.

Detailed comments are provided in the Appendix.

6. Appendix

Issue	Recommended Change	Rationale
Lake Selection		
Potential for eutrophication in study lakes; how representative are lakes	Confirm that it is change in chemistry of acid sensitive lakes that is of interest and acknowledge that this does not likely reflect the response of the lakes in the region. Consider N as a stressor in its own right, not just as a component of the PAI	Management decisions depend on what is being assessed: a negative change in a chemical endpoint in 15% of ASL in a sub-region may only correspond to 2% of all lakes – is this OK? N emissions in the region may exceed S and so eutrophication effects should be considered. The appropriate chemical measures are made, but the current lakes may or may not be appropriate.
Potential for episodic acidification	Evaluate whether the current seasonal sampling captures snowmelt in the 10 ASL. If yes – compare the chemistry of this sample with other seasons. If no, do a spring survey of selected ASL.	For many Boreal systems, the spring snow melt represents the largest influx of water to most lakes and has been associated with episodic acidification (Laudon et al. 2004). This is not currently assessed in the ASL.

Between-year comparisons of measurement endpoints over the entire population of lakes	This analysis is really only for descriptive purposes and should not be used for assessing potential impacts of acid deposition.	Even if these analysis were done for each sub-region rather than the entire data set, between lake variability in chemical parameters is so great that extremely large changes would be needed to cause a significant change. As all lakes will not respond to the same extent there is a large danger of biologically meaningful changes in some lakes not being identified. In any case a repeated measure ANOVA should be used.
Critical Load Calculation	Clarify why this is done and review the methodology following comments below.	During EIAs conducted in the oil sands region, ASL are almost entirely evaluated using a critical loads approach. While steady state critical loads are calculated using the SSWC model, they do not appear to be used for assessment given the large number of lakes that exceed the

		calculated critical load – yet negligible effects are reported in the 2009 annual report. Presumably this is because considers exceedance of the CL to signify a potential effect, not a real one. If this is the case – why do them? By definition, steady state critical loads should not change over time. They should not be calculated annually – rather they should use the average chemistry from multiple years.
BC_o	Justify the rationale for assuming BC_o (pre-industrial base cation concentrations) are the same as currently observed.	In all other applications of the SSWC, BC_o is estimated by taking into account the fact that current base cation concentrations are elevated due to acid inputs. This is done by estimating the increase in sulphate above pre-industrial levels and assuming all nitrate is of anthropogenic origin. The

		F-factor is used to estimate the increase in base cation concentration associated with increase in acid input. Therefore in the approach used by RAMP it is implicit that there has been no increase in S due to industrial activities, which I find surprising [note the lack of correlation between H and sulphate is not support for this as this in only expected in acid sensitive lakes receiving substantial acid inputs – I would not classify many of the lakes as acid sensitive]. If it is assumed that there is no increase in lake sulphate, but measurements and models indicate that there has been an increase in S deposition, then one of the assumptions of the

		SSWC is violated as the model assumes conservative behaviour of S (i.e. S deposited goes into the lake). If it is confirmed that there has been no increase in S deposition in the region, then the estimation of BC_o is correct – however estimating exceedance using PAI is then incorrect as base cations are being double counted (i.e. current lake base cation concentrations include both catchment and deposition sources. Using PAI (S+N-BC) effectively counts base cations twice and therefore reduces the estimated exceedance of the critical load. If RAMP cannot confirm that S deposition has not increased then the increase in sulphate (due to S deposition) in lakes

		needs to be accounted for – which is problematic for two reasons: 1. S behaviour in catchments is likely not conservative due to the large proportion of wetlands (Whitfield et al. 2010), which reduces S and immobilizes it within the catchment and therefore does not contribute to acidity. 2. Some of the lakes (in the Birch Mountains) clearly have a natural source of S, which would need to be accounted for in the estimation of pre-industrial sulphate
ANC_{lim}	Justify the Rationale for this limit and change if necessary	The SSWC uses a critical chemical criteria that usually ranges from 0 – 40 µeq/L, yet RAMP uses a highly conservative 75 µeq/L (based on discussions in WRS, 2004). I have not seen this report, but my feeling was that a high value was selected because of

		the high natural organic acidity. This would generally be OK – but not for all lakes as the calculation of a negative critical load effectively indicates that an ANC of 75 µeq/L is not achievable (*i.e.* critical limit is too high). Secondly, the SSWC equation used by RAMP now also includes the impact of organic acidity on the critical load (see below) – so again this indicates that the acidic influence of organic acidity is being counted twice, which results in a very conservative (low) critical load – and is the reason why negative values are obtained.
ANCorg-A-$_{SA}$	Recalculate the impact of organic acidity following Lydersen et al. (2004)	It has recently been recognised that in regions with high DOC, the role of organic acids needs to be considered when calculating the critical load. In this respect the approach

		described in RAMP (2005a) [sometimes referred to as RAMP (2005)] is correct; however the formula they use is problematic for a steady state model. The calculation of ANC_{org} = 0.0068 * DOC * exp (0.8833*pH) includes pH, which is a dynamic term in lake acidification. If the calculation is performed using today's pH measurement it is incorrect, because at critical load (ANC 75µeq/L) the same lake will have a different pH – which will give a different ANC_{org} and hence a different critical load. The whole point of a steady state critical load model such as the SSWC. I suggest RAMP uses the values suggested by Lydersen et al. (2004) and used in Jeffries et al. (2010) – and that this term be incorporated into the ANC_{limit} (i.e. ANC_{limit} =

		value (probably not 75) + (10.2/3) * DOC (mg/L)) so that organic acidity is not counted twice.
Runoff	Choose one value – I recommend the lake specific values from isotopes.	The calculation of the critical load includes many assumptions. Runoff is one of them. I would suggest using just one runoff value and that is the lake specific value from isotope measurements. Using two values just adds confusion.
Negative Critical Loads	Evaluate whether the SSWC approach is appropriate for these lakes.	As mentioned above – the calculation of a negative critical load means the calculation is incorrect (due to the choice of a high ANC_{limit} (plus in this case, the additional contribution of natural organic acidity).
PAI	PAI should not be used for estimating critical load exceedance.	I make this suggestion for 3 reasons. 1. As stated above, using PAI effectively counts the buffering effect of base cations twice (it is a component of lake chemistry (BC_o).

		2. In most cases SSWC is used to assess sensitivity to S deposition as it is recognised that N does not behave conservatively in the catchment. Most other applications use the currently observed lake nitrate concentrations and assume no further increase in lake nitrate. If there are concerns over the potential increase in lake nitrate then another model (FAB-First order acidity balance – see Aherne et al. 2004 as an example) should be used – and this would need additional data (i.e. proportion of catchment with wetlands). Using PAI, effectively is a worst case scenario and assumes all N is leached as nitrate into the lake – which is not likely. 3. Exceedance is usually calculated assuming S is conservative in the catchment. To repeat my

		earlier comments – if S deposition has increased in the region, but this is not reflected in lakes then this assumption is violated. Finally PAI is not useful as I have no idea how this value relates to actual deposition – when does potential become reality?.
Shewart Charts	Remove these	I do not feel that the use of Shewart charts is particularly useful. In several cases a change of 2 or 3 SD would be extreme and far beyond the level where damage could occur. They should not be used for assessment purposes
Data Errors	Check database and reports for errors in units	In the database the numbers are often presented to 5 or 6 significant figures, which implies a level of precision that is not achievable. SBC is presented as meq/L, when it should be µeq/L.

		It would be good to clarify for parameters should as nitrate, whether the value is for nitrate or nitrate-N (they are different). In the RAMP 2009 report, nitrate values are reported as mg/L when they should be µeq/L (i.e. Table 7.5-1. There may well be others that I have not caught and suggest someone goes through this very carefully.
Inconsistencies in the text	Similar to the database and units – it would be beneficial to check the text for consistencies or errors.	In the Technical Design and Rationale Report, impacts on biology are reported as an objective (3-144). There is no biology data in RAMP and I am not sure whether any attempt has been made to relate chemistry to biological effects. In the 2009 Ramp report summary – 'A statistically significant change in any of the measurement endpoints beyond natural variability, resulting in a reduction of lake pH, Gran alkalinity,

		Critical Load or base cation concentrations or an increase in nitrates or aluminum concentrations'. Reduction in critical load or base cations is incorrect and sulphate is not mentioned.

7. References

Aherne, J., Posch, M., Dillon, P.J., Henriksen, A. (2004). Critical loads of acidity for surface waters in south-central Ontario, Canada: regional application of the first-order acidity balance (FAB) model. Water Air and Soil Pollution: Focus, 4: 25-36.

Burns, D.A., McHale, M.R., Driscoll, C.T., Roy, C.M. (2006). Response of surface water chemistry to reduced levels of acid precipitation: comparison of trends in two regions of New York, USA. Hydrological Processes, 20: 1611-1627.

Davies, J.J.L., Jenkins, A., Monteith, D.T., Evans, C.D., Cooper, D.M., (2005). Trends in surface water chemistry of acidified UK freshwaters, 1988-2002. Environmental Pollution, 137: 27-39.

Evans, C.D. (+ 9 others) (2001). Recovery from acidification in European surface waters. . Hydrology and Earth System Sciences, 5: 283-297.

Jeffries, D.S., Semkin, R.G., Gibson, J.J., Wong, I. (2010). Recently surveyed lakes in northern Manitoba and Saskatchewan, Canada: characteristics and critical loads of acidity. Journal of Limnology, 69 (supp 1): 45-55.

Henriksen, A., Dillon, P.J., Aherne, J. (2002). Critical loads of acidity for surface waters in south-central Ontario, Canada: regional application of the steady state water chemistry (SSWC) model. Canadian Journal of Fisheries and Aquatic Sciences, 59: 1287-1295.

Laudon, H., Westling, O., Bergquist, A., Bishop, K., (2004). Episodic acidification in northern Sweden: a regional assessment of the anthropogenic component. Journal of Hydrology, 297: 162-173.

Lydersen, E., Larssen, T., Fjeld, E. (2004). The influence of total organic carbon (TOC) on the relationship between acid neutralizing capacity (ANC) and fish status in Norwegian lakes. Science of the Total Environment, 326: 63-69.

Rapp, L., Bishop, K. (2009). Surface water acidification and critical loads: exploring the F-factor. Hydrology and Earth Systems Science Discussions, 6: 3917-3945.

Skjelkvale, B.L., Mannio, J., Wilander, A., Andersen, T., (2001). Recovery from acidification of lakes in Finland, Norway and Sweden, 1990-1999. Hydrology and Earth System Sciences, 5: 327-337.

Stoddard, J.L. (+22 others) (1999). Regional trends in aquatic recovery from acidification in North America and Europe. Nature, 401: 575-578.

Watmough, S.A., Aherne, J., Dillon, P.J. (2004). Effect of declining base cation concentrations on freshwater critical load calculations. Environmental Science and Technology, 39: 3255-3260.

Whitfield, C.J., Aherne, J., Gibson, J.J., Seabert, T.A., Watmough, S.A. (2010). The controls on Boreal peatland surface water chemistry in northern Alberta, Canada. Hydrological processes. DOI: 10.1002/hyp.7637