Final Report

ESSA 35 YEARS

Environmental Monitoring Triggers

Workshop Report

Prepared for Kelly Munkittrick, Canada's Oil Sands Innovation Alliance

Environmental Monitoring Triggers

Workshop Report

Prepared for:
Kelly Munkittrick
Canada's Oil Sands Innovation Alliance (COSIA)

Lead Authors:
Lorne Greig and Darcy Pickard

Contact:
Lorne Greig
lgreig@essa.com
905-770-6334

Darcy Pickard
dpickard@essa.com
604-733-2996

Suggested Citation:
Greig, L. and D. Pickard. 2014. Environmental Monitoring Triggers: Workshop Report. Prepared for Canada's Oil Sands Innovation Alliance, Calgary, Alberta. 27 pp + appendices.

Cover Photo:
Kananaskis (Photo by: Darcy Pickard)

ESSA Technologies Ltd.
Vancouver, BC Canada V6H 3H4
www.essa.com

Acknowledgements

We extend our sincere thanks to all of the workshop participants for their thoughtful contributions to the workshop discussions. We thank Tim Arciszewski, Tim Barrett, Simon Courtenay, Allan Curry, Heather Keith, Bruce Kilgour, Chantelle Leidl, Brian Pyper, Kathy Racher, Jim Schieck, Keith Somers, Raju Neal Tanna, and Kim Westcott for their valuable presentations during the workshop. Special thanks to Kelly Munkittrick for conceiving and enabling the workshop, to Meghan McEvoy for her work organizing the workshop logistics and to Tim Arciszewski, Tim Barrett, Bruce Kilgour, Brian Pyper, and Keith Somers for their analyses and discussion of the sample GSI dataset.

Table of Contents

List of Figures

List of Tables

1 Introduction

This report presents the results of a two-day workshop to evaluate alternative approaches and identify important considerations when developing environmental monitoring triggers. Groups invited include those conducting studies dealing with the Athabasca River, Bow River (Alberta), metal mines (BC), northern mines (ON, NWT and Nunavut), pulp mills and hydroelectric development (ON), hydroelectric development (NB), sediment contamination (NB), and some provincial and national monitoring programs. This workshop held in Kananaskis, Alberta, June 2014, provided the opportunity for scientists to share ideas, questions, and lessons learned from a variety of perspectives.

1.1 Background

Recent years have seen increased interest in long-term regional monitoring programs while, simultaneously, many programs have been drastically reduced or cut as cost-savings measures. As well, there is increasing appreciation for adaptive monitoring frameworks, which can efficiently focus resources and maximize information for decision-making. The Canadian National Environmental Effects Monitoring programs for pulp and paper mills and for metal mines are examples of long term, cyclical, adaptive monitoring programs. Considerable effort went into defining monitoring cycles and the development of triggers to guide site-specific evolution of programs through a series of different phases.

Monitoring programs need to consider three types of protection: protection from pursuing an effect that is real, but small and meaningless, protection from “finding” an effect that is not real (false positive - type I statistical error) and protection from missing a real effect (false negative - type II statistical error).

The power to detect an effect in a long-term monitoring program increases with increasing data, and with sufficient data it becomes possible to detect small differences that are statistically significant, even though they may be biologically or ecologically meaningless. Protection against this concern is dealt with in some programs by defining a “critical effect size”, which represents either the size of a difference that is considered meaningful, or the size of a difference that a program is designed to be able to detect. There are multiple ways programs have tried to define this critical effect size (Munkittrick et al. 2009).

There is also the chance that these small differences may reflect natural variability and not a real difference. Variability in estimates of an endpoint caused by things other than the influence of interest constitutes “noise” muffling out the signal of interest. Noise in data originates from a variety of sources, including natural variability, the adequacy of reference sites, the ecological relevance of the endpoint, the presence of confounding factors, sampling error and statistical error. Characterizing this noise is crucial to avoid investing significant effort in false signals. With high natural variability, very high replication may be required for signal detection and this may be prohibitively expensive and/or may require an unacceptably long time to be a useful to trigger action in an adaptive monitoring program.

In either case, an adaptive monitoring program should respond to an effect detection (a change greater than some critical effect size) by triggering confirmation steps which include changing the frequency of monitoring, and potentially the extent and magnitude of sampling. If such steps fail to confirm the effect, the program can revert to its previous monitoring state and intensity, and there is enhanced protection from any further false positives. False negatives, on the other hand, are a real concern because over time persistent changes that go undetected have the potential to translate into broader and more significant impacts.

In any adaptive monitoring program, the consequence of not detecting an effect is to keep monitoring at the existing level, or after some confidence builds, to reduce monitoring intensity and frequency. The only real protection from false negatives is increased power and the driving force affecting power is the variability of the measurement endpoint. Other factors affecting power (critical effect size, Type I and Type II error levels) can be adjusted but the greatest challenge in designing a sensitive monitoring program is to reduce endpoint variability.

The complex tradeoff in any monitoring program is to ensure focus is on meaningful changes and to avoid chasing “ghost signals”, while still ensuring true effects aren’t missed. The end result is that an effective adaptive monitoring strategy needs to ensure sufficient power or set conservative triggers that protect against false negatives (Type II errors), and that protection against false positives (Type I errors) is achieved by confirmation monitoring. There are more than a dozen cumulative effects monitoring programs across Canada which are in the process of developing triggers to adapt monitoring, and there are a variety of approaches and philosophies being considered within those programs.

1.2 Workshop Objectives and Workshop Approach

The objective of this workshop was to develop a series of principles to be considered when triggers are developed. We endeavored to do this through:

- Developing a common understanding of the various approaches currently in use to develop triggers and evaluate the strengths and weaknesses of these approaches; and
- Developing a common terminology around monitoring triggers.

1.3 Workshop Format and Participation

To provide a common understanding of the purpose, scope, and approach to the workshop, discussions began with an overview of the workshop objectives and agenda. This was followed by an introductory presentation summarizing key terminology and several case studies which illustrated the need for better guidance and consistency in approaches to setting monitoring triggers.

The majority of the first day was allocated to 11 invited presentations describing current approaches for setting triggers in different case studies from across Canada (Figure 1). In addition, each presenter was asked to apply their approach to an example dataset (consisting of 16 years of female gonad size data for a site exposed to pulp mill effluent and its reference site) provided prior to the workshop. Short discussion sessions held every 3 or 4 presentations provided the opportunity to ask questions, clarify terminology and identify emergent themes. The workshop discussions closed with a round table review of key insights and outstanding issues the workshop participants hoped to tackle further during the workshop.

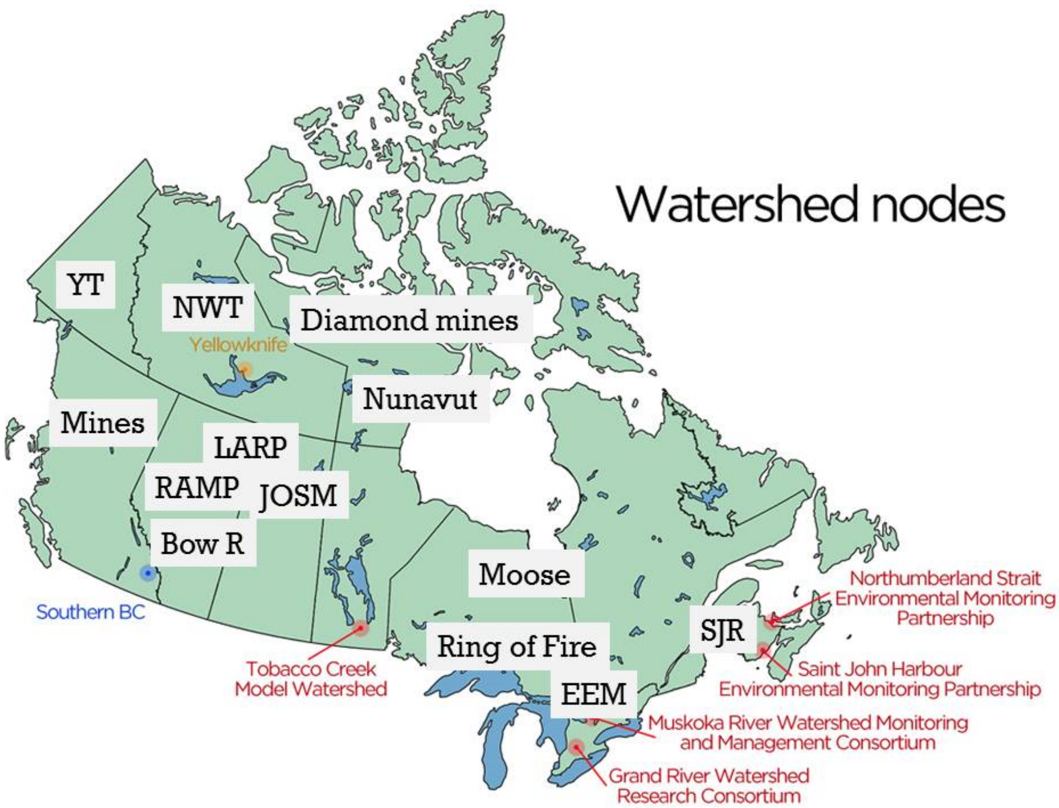


Figure 1. Locations of the various case studies presented during the workshop or represented by participants at the workshop.

The second day of the workshop was originally scheduled to focus on a data analysis exercise using two additional example datasets. However, it was determined that the female gonad size example presented on Day 1 had provided sufficient insight into the different approaches and resultant outcomes. Instead, the group decided to focus on the most common outstanding issues identified at the close of the previous day.

The detailed workshop agenda is given in Appendix A, and the workshop participant list is provided in Appendix B.

2 Summary of Current Approaches

Workshop presenters were asked to provide a short (15 minute) presentation demonstrating their approach and experiences developing and applying environmental monitoring triggers. In particular they were asked to: describe the context of their study; how they defined normal; their approach to developing triggers along with an example; their approach to selecting reference sites if applicable; whether or not they used a tiered or adaptive monitoring approach; and any issues or challenges they would like to discuss. Table 1 provides a high level summary of these presentations focusing on the components that emerged as the most important to inform principles for developing triggers (i.e., context, question definition, approach, and the tiers or levels of triggers).

Table 1 High level overview of the workshop presentations.

Context	Questions	Approach	Tiered?
Moose River pulp mill closure in 2006. 6 years pre-data, 3 years post. Upstream and downstream of mill.	1) Δ within site over time. 2) difference between sites over time. 3) compared to regional reference.	>2 SD of 3 year grand mean Error rates (false positives early, false negatives later) Overlap of error bars Pulse vs press disturbance	Yes. EEM style (e.g., confirmation; E&M)
St. John Harbor. Extensive activity: Oil refinery, pulp and paper mills, brewery, active port, municipal wastewaters, dredging	1) Characterize Baseline Concentrations. 2) Characterize hotspots	>3 SD of mean across all sites/years to date Unclear if they should separate inner and outer harbour results/triggers Use all data to calculate trigger or more recent data only? Given low temporal variance, how often do we need to resample sediments	No
Benthos monitoring in the oil sands	1) Trends over time/comparison to baseline (i.e., BACI). 2) Comparison of reach annual means to normal ranges of variation. 3)	Comparisons to normal based on testing for a difference of at least 2SD (via equivalence testing approach) Normal for the test reach, based on a minimum of 3 years data. Regional normal (RCA approach) Temporal variations in baseline reach data are not often	Yes. Basic; Confirmation; Extent; Cause; Concern

	Indicator species approach	correlated with measurable climatic or discharge events, causing drift in normal ranges, and potential inflation of noise Tough to get sufficient baseline data. Spatial variation makes it difficult to find appropriate reference sites. Reference sites may switch to impacted sites over time.	
Meadowbank Mine, NU. Impacts to water quality, fish habitat and fish populations.	BACI comparison of nearfield, midfield, far field stations to 2 reference lakes	Basin scale triggers Normal = Pooled reference/pre-exposure data (all years/sites, medians and 95th percentiles) Where thresholds exist these are used if no threshold, 95th percentiles are used. 72 variables! Limited before data	Yes.
Snap Lake Diamond Mine. Plankton data: 9 yrs. Multiple exposure and multiple reference stations. Fish tissue chemistry: 4 yrs. 1 exposure and 2 reference lakes	Significant difference between exposed and reference site. If Yes, see if exposed is outside of the range of normal.	95% Prediction Interval for a single observation when μ and σ are unknown. Considers variability in the estimate of μ and σ. We test for normality and transform data as appropriate Tiers? Confirmation, E&M Should we use the annual means or the raw data?	No
Surface water quality in the lower Athabasca River	Comparison to static regional reference based on historical data.	38 WQ indicators, triggers for each indicator, limits for 21 • Mean triggers are the arithmetic mean of the historical data. Peak triggers are the 95th percentile of the historical data. • Normal vs noise is defined as a statistically significant change from the historical data. • Static triggers Prone to Type I errors Requires extensive time-series Annual assessment/response can be resource intensive.	Yes. Triggers, Limits, and Management Response levels.
Bow River Central Reach. Industrial facility closed in 1992. GW treatment since 2002.	Δ within site over 1) Δ within site over time. 2) difference between sites over time. 3) compared to regional reference.	Statistical difference between sites or 90 or 95th percentiles for within site thresholds. Normal = Historical data 2002-2013 Changes to river morphology	Yes. Confirmation; Action; Concern

		Changes undermining validity of reference station(s) Changes in municipal and industrial loading	
Alberta Biodiversity Monitoring Institute For each species determine how abundance varies: among habitats & in relation to human disturbance	Compare estimates of degraded areas to pristine or snapshot in time.	Use models relating human footprint to abundance to determine thresholds and natural variability. For reclamation purposes, more interested in the trajectory than the current condition. In practice, society dictates the level of disturbance that is considered acceptable.	No
Groundwater Monitoring, North Athabasca Oil Sands Established in 2009. 46 wells, 11 sites (5 years)	Detect departure from natural conditions	Interim Triggers: 75th percentile. Normal for each site is based on (first 8 years) mean +/-4.5 SD. Trends assessed but not a formal trigger. Outliers generally discarded. High spatial variability Cost/infrastructure	Yes. Interim & baseline triggers, plus 7 management response levels.
Biological condition of Ontario's inland lakes and streams	Compared to local or regional reference	All use mean +/- 2 SD's to define normal for set of appropriate lakes. OBBN & FIRRNO – use 'best' subset of (700+; 160 respectively) available sites as reference Inland lakes: 16 reference lakes, 2 time periods. Single sample in time Large spatial range – how to select 'best' reference sites Noise – e.g., seasonal variation, within site variation. Not yet linked to management actions Understanding relationship between each metric with increasing concentration of common stressors? Cumulative effects – how do we identify/quantify the separate effects from multiple stressors acting together?	Possible. Interim Trigger based on percentile position vs. statistical significance.
NWT. Absence of federal regulations for diamond mines	Comparison of nearfield, midfield, far field stations to control stations	 Significance thresholds (limit of acceptable change we never want to hit these, value based decision) Site-specific Action Levels: • exceedances of normal range and/or reference	Yes. L, M, H action levels.

		conditions • guideline values if appropriate (e.g., water quality, sediment, fish tissue) • Importance of water body to stakeholders • Traditional knowledge end points (e.g., taste) Erring on the side of a "false positive" is of low risk because there are several action levels between baseline and significance threshold. Also, each exceedance is evaluated before substantial action taken. It is hard to define a limit of acceptable change (significance threshold)	
Mactaquac Aquatic Ecosystem Study. Dam is at the end of its life and options to remove or rebuild are being considered. 4 years to make a decision. Construction in 8 years	Comparison to itself over time.	Not yet developed	N/A

3 Data Analysis Exercise: Female Gonad Size

An example data set shared with workshop participants prior to the workshop provided the opportunity to evaluate alternative approaches to defining triggers on a common dataset. The data consisted of gonad size data as a percentage of body weight (GSI) for a reference (R) and exposed (E) site. Twenty replicates were provided for each site in each of 16 years. The example was based on real data but fictionalized to provide an example to put the various approaches in a similar context for discussion. Five participants attempted the exercise and each used a slightly different approach resulting in different conclusions (Table 2; Figure 2).

Table 2 Summary of the approaches and conclusions of participants for the data analysis exercise.

Presentation	Summary of approach	Conclusion(s)
1	Compared each site to itself over time. Normal1 was defined as historical mean +/- 2 SD, where normal was updated as new data was obtained (so long as the new point was not considered unusual). Trigger = +/- 2 SD of mean.	Possible regional effect in years 6-8. Effect observed in last 3 years at exposed site. Figure 2 (i)
3	Compares both the Reference (R) and Exposed (E) sites to normal range estimates (based on historical data) for both reference and exposed sites over time. While 2 SD was the threshold, a 95% tolerance region was applied to the estimate for each SD value, this results in a smaller window before a trigger is reached.	E site normal: R almost always exceeds the trigger and E exceeds it in the last 3 years. R site normal: R exceeds in years 6-8. E exceeds (on lower end) in most years excluding the last 3. Figure 2 (ii)
4	Used BACI approach to look for temporal change in Reference site relative to Exposed site.	Depending on trigger, effects detected in years 10-12 as well as 14-16. Figure 2 (iii)
10	Compared the Exposed site to the absolute values at the Reference site over time. Used 2 SD based on year 1 at the Reference site to define normal.	Found that in most years E was well outside of normal for R. In the last 3 years, E comes into the normal range. Figure 2 (iv)
5	Test for significant difference between Reference and Exposure sites. If significant then does the magnitude of difference lie within the range of natural variability? Where normal is defined by the Reference site.	N/A

[1] In general, throughout the workshop, the term 'normal' was used to describe the desired condition, with deviations from normal being a cause for concern.

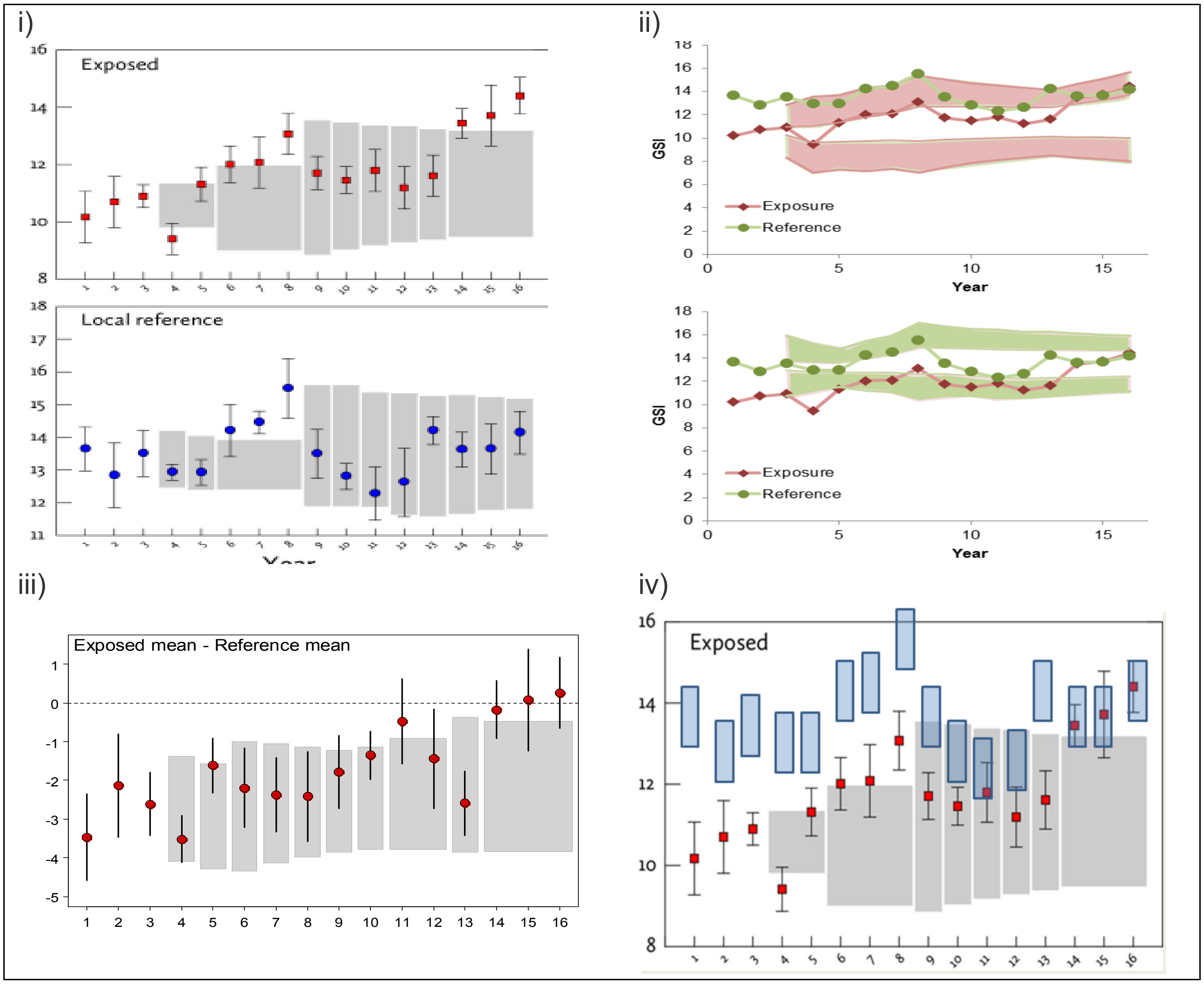


Figure 2 Outcomes of 4 different approaches to applying triggers to the example female gonad size dataset.

The data analysis exercise was revisited on the morning of the second day. While there were inconsistencies in the approaches and conclusions, these were primarily due to differences in the questions each participant asked and the way in which they defined what condition was considered to be normal. Further discussion found that each approach was valid for the question it was asking and in fact there was general acceptance that 2 standard deviations was an appropriate trigger for defining normal. There were however outstanding uncertainties in determining what data should be included in the definition of normal and whether or not this should be updated over time.

4 Outstanding Issues

At the end of the first afternoon, workshop participants identified a variety of outstanding issues that were grouped into six priority themes for further discussion:

- What are the questions?
- How is normal defined?
- Methods/Approaches (Data availability, Statistical approaches, Power)
- How / when should values be incorporated?
- Accounting for complexity of ecological systems
- Process/management frameworks

The detailed notes from this discussion are provided in Appendix C. This section provides a summary of the subsequent discussions for each theme held during the second day of the workshop, more detail is provided in Appendix D.

4.1 What are the questions?

During the course of the workshop, it became evident that the context and questions of interest were the most important factors in determining the approach for developing triggers. At first glance, each of the five participants who attempted to apply their approach to the example dataset produced very different triggers and results. However, upon further review (Section 3) the participants realized that the key difference was in the question asked. Therefore, determining the most common set of questions is a critical first step to providing general guidance for setting monitoring triggers.

Two workshop participants: Bruce Kilgour and Keith Somers presented a scenario (Figure 3) that helped the group to better communicate their questions of interest and outcomes of concern. In this figure, the small orange curves represent the distributions within a lake (e.g., across years or samples), the large blue curve represents the distribution of regional reference lakes, and the large brown curve represents the distribution of regional impacted lakes. Some might be interested in whether or not their lake, which started off with a very abundant fish population, shows a decline, even if it is still within the regional range of normal abundance. Others might only be interested in whether or not it is outside the regional normal. Still others might be interested in how it compares to other impacted lakes. The triggers will differ dramatically for each of these questions.

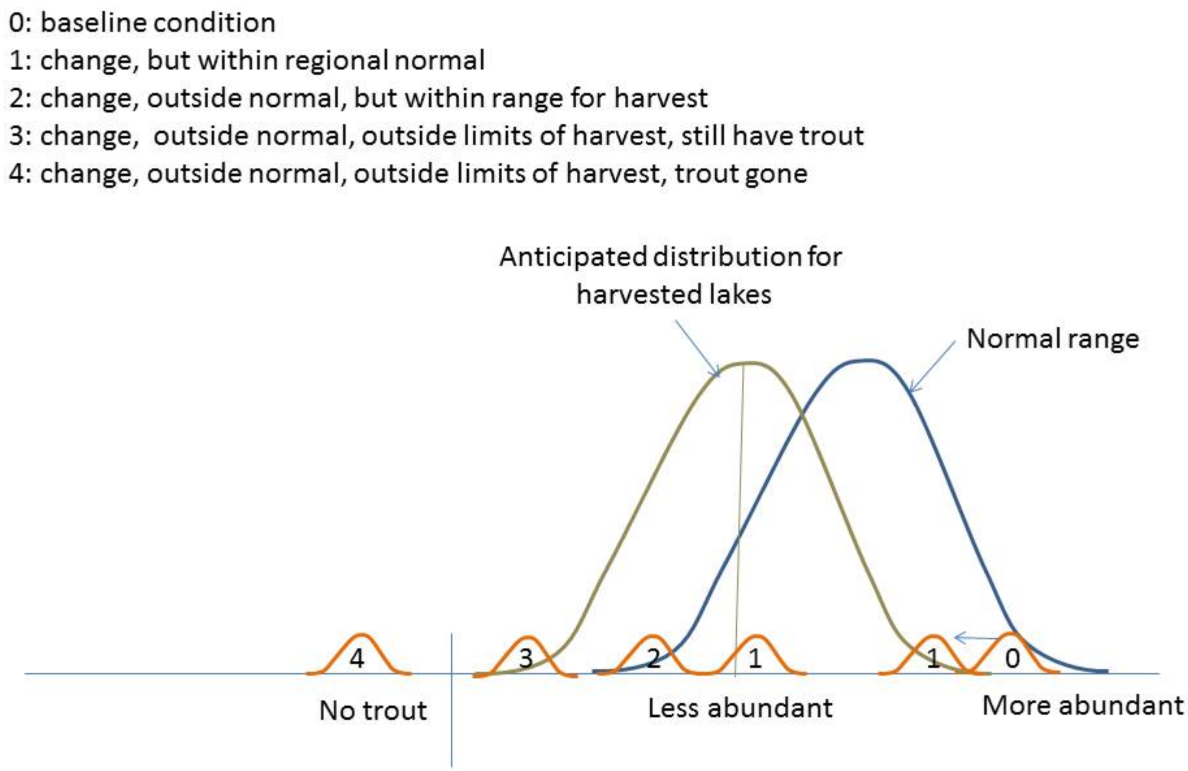


Figure 3. A simple scenario for the purpose of clarifying the potential questions of interest. The small orange curves represent the distributions of fish abundance within a lake (e.g., across years or samples), the large blue curve represents the distribution of regional reference lakes, and the large brown curve represents the distribution of regional impacted lakes. 0 represents a hypothetical baseline and 1-4 represent hypothetical shifts from the baseline after harvest.

Using this figure to ground the conversation the workshop participants reached agreement upon five common questions of interest (Table 3). The order refers to the priority order identified by the workshop participants. There was some discussion about whether or not the order would change depending on the availability of historical data, but in the end the same order was identified for both scenarios.

Table 3 Common environmental monitoring questions of interest.

Question	Order	Description
Is my site different from local reference sites?	1	A comparison of absolute values (e.g., abundances) between the site of interest and local reference sites.
Has my site changed (relative to itself)?	2	A comparison over time at a given site.
Has my site changed relative to local reference sites?	2	Comparison of relative changes between the site of interest and local reference sites. Are they tracking each other?
Is there a change relative to regional reference?	2	A comparison of absolute values between the site of interest and regional reference sites.
Where is my site relative to other impacted sites?	3	A comparison of absolute values between the site of interest and regional impacted sites.

Brian Pyper then proposed an organizing structure (Table 4) illustrating how the analytical approaches might vary by question depending on the following two axes: 1) spatial scale (site, local, regional); and 2) the comparison type (absolute vs change). This organizing structure serves as the beginning of an analytical framework providing guidance to practitioners on how to approach development of triggers depending on the context and question(s) of interest. Further work is required to refine this guidance. Appendix E provides some additional post-workshop thoughts on how this table might be used.

Table 4. Organizing structure illustrating how the analytical approaches for monitoring triggers may change depending on the data availability, spatial scale, and comparison type for a given question.

	Comparison type	
Spatial context	Absolute comparison	Temporal change
Exposure site only	(1) Not applicable	(2) EEM analogy: Before-after (BA) design Analysis: BA t-test, trend analysis, or intervention analysis (step/pulse)
Paired exposure and (local) reference site	(3) EEM analogy: Control-Impact (CI) design Analysis: CI t-test	(4) EEM analogy: Paired BACI (BACIP) design Analysis: Paired BACI t-test, paired trend or intervention analysis (step/pulse)
Multiple reference sites (local/regional) [extends to multiple exposure sites as well]	(5) EEM analogy: Multiple Control-Impact (MC-I) design; reference condition approach (RCA) Analysis: ANOVA, multivariate time-series models, ordination, etc.	(6) EEM analogy: Multiple controls BACI design (Underwood 1994) Analysis: ANOVA, multivariate time-series models, etc.

4.2 How is normal defined?

Triggers for all of the questions described above require us to define a point of concern, typically a departure from normal. However, there was a great deal of variety in how different groups defined normal and the approaches were highly dependent on the data availability and the question of interest. In an ideal world, participants agreed that normal would be defined by appropriately selected reference sites and/or pre-exposure data at the impact site depending on the question of interest. In many cases there are limited choices for reference sites and there may be limited or no pre-exposure data. Even when data are available there are a number of outstanding considerations.

The reference condition approach is very sensitive to how well you have matched your reference sites to your treatment site. People often lump things that shouldn't be lumped. For example, if you include both fine and coarse sediment lakes in your definition of normal, you may not detect the fact that you have moved beyond normal for a coarse sediment lake because you are still within the broader regional range. It is very important to have rigorously defined comparable lakes (or other experimental units of interest).

Several participants raised the point that the concept of 'pristine' or 'natural' may not be a realistic comparison. It is very difficult to find a reference site that isn't impacted by something (e.g., hydro, logging, roads etc.). Minimally disturbed might be a more realistic comparison.

Likewise there is no simple answer to how normal should be redefined temporally. How should the baseline temporal window be determined? Should normal be updated over time and if so how often? Does this depend on whether or not pre-exposure data is available? For example, Environment Canada uses a 30 year moving window to define normal for climate variables. What if a reference site begins to drift, should it be removed from the reference population? There was some disagreement among participants about whether or not the definition of normal should be updated over time. Normal depends on perspective: permitting timelines, career timelines, lifetimes, multiple generations etc. However, it is important not to be too relaxed with the rules on defining normal as it might lead to situations where you simply change the definition of normal when you get an undesirable result.

In general, participants agreed that if reference sites and/or pre-exposure data were available these should be used to define normal. When no pre-exposure data are available, the definition of normal should be updated annually until it stabilizes, or alternatively spatial replicates might be used in the early years. Different definitions of normal may be used for different questions recognizing that you don't need to pick just one. It may also be helpful to evaluate several questions at the same time (e.g., absolute comparison to a near pristine benchmark as well as incremental changes relative to reference sites). Depending on the data availability some questions may not be able to be addressed, more discussion /

research is necessary to provide suitable guidance for the best approach under different data availability scenarios.

4.3 Methods/Approaches

4.3.1 What is an environmental monitoring trigger?

"Trigger: Something that causes something else to happen", Merriam-Webster

Workshop presentations illustrated a broad use of the term trigger and varied approaches for defining triggers across programs. Initially the term trigger and critical effect size were used interchangeably. However, later discussions clearly indicated that there was also a need for early warning triggers that could prevent a critical effect being reached. Participants agreed that there could be and probably should be multiple triggers corresponding to reasonable size steps and different action levels which can easily be communicated. The following levels were proposed:

Don't worry be happy No action required
This is weird Low action level, this must be well within the 'Freak out levels'. Exceeding this level might trigger further data collection.
This is weird and abnormal Medium action level (could be a critical effect size), exceeding this level might result in a study to determine the extent and cause of the effect.
Freak out level. High action level. This corresponds to the level at which ecologically relevant changes (could be a critical effect size) or changes which impact stakeholder values (e.g., are the fish safe to eat?) occur.
PNR. Refers the point of no return, or very costly return.

4.3.2 Approaches for defining triggers

A primary focus of this workshop was to explore the ways triggers may be defined. Triggers might be defined based on biology, values, legal requirements, or based on statistical differences from normal, however normal is defined. This section focuses on the latter case.

All of the groups that presented in the workshop defined their triggers in terms of the mean for a single indicator. The most frequent trigger used was 2 standard deviations (SD) from the mean. However, 3 and 4.5 standard deviations were each used in one example. Another approach presented was to test whether or not 2 SD was exceeded, rather than testing for a difference from zero (equivalence testing, Kilgour et al. 1998). Conceptually imagine a confidence interval on the trigger itself (Figure 4). The choice of 2, 3, 4.5 standard deviations depends on the risk tolerance of the given program and the trigger level (e.g., low action vs high action), each is associated with a different probability of making a Type I error (e.g. 0.05, 0.01, $6.8*10^{-6}$). For example, with a random sample from a stable normal distribution, we'd expect 95% of the observations to lie within +/- 1.96 standard deviations of the mean, and 5% to exceed this value. The probability is likely a better way to communicate the choice of triggers than to use the standard deviations. With small sample sizes, the estimate of standard deviation may not be very accurate and a t-distribution rather than normal should be employed to calculate the multiplier (e.g., 1.96) that corresponds to the 95% confidence level. Alternatively the equivalence testing approach will account for the uncertainty in the trigger. The estimate of variance (and standard deviation) should stabilize with 8-16 replicates (spatial or temporal) assuming the distribution has not changed and therefore with large sample sizes these approaches will converge.

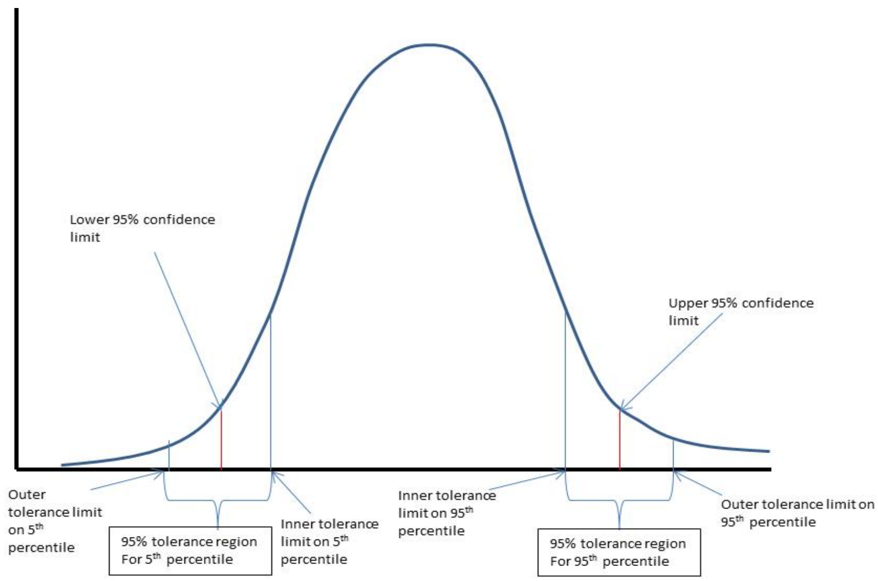


Figure 4 Difference between simply using the +/- 1.96 standard deviations (red line) as the trigger and accounting for uncertainty in the trigger itself by applying tolerance regions. As sample size (n) increases, the width of the tolerance regions will shrink.

4.3.3 Early warning triggers

Participants were generally comfortable that 2 standard deviations was a reasonable medium or high action level trigger in absence of other information. However, there was also substantial discussion about additional "early warning" or low action triggers. Figure 5 illustrates a number of patterns in data that might indicate that further investigation is warranted even though a high or unacceptable level action level trigger has not yet been exceeded.

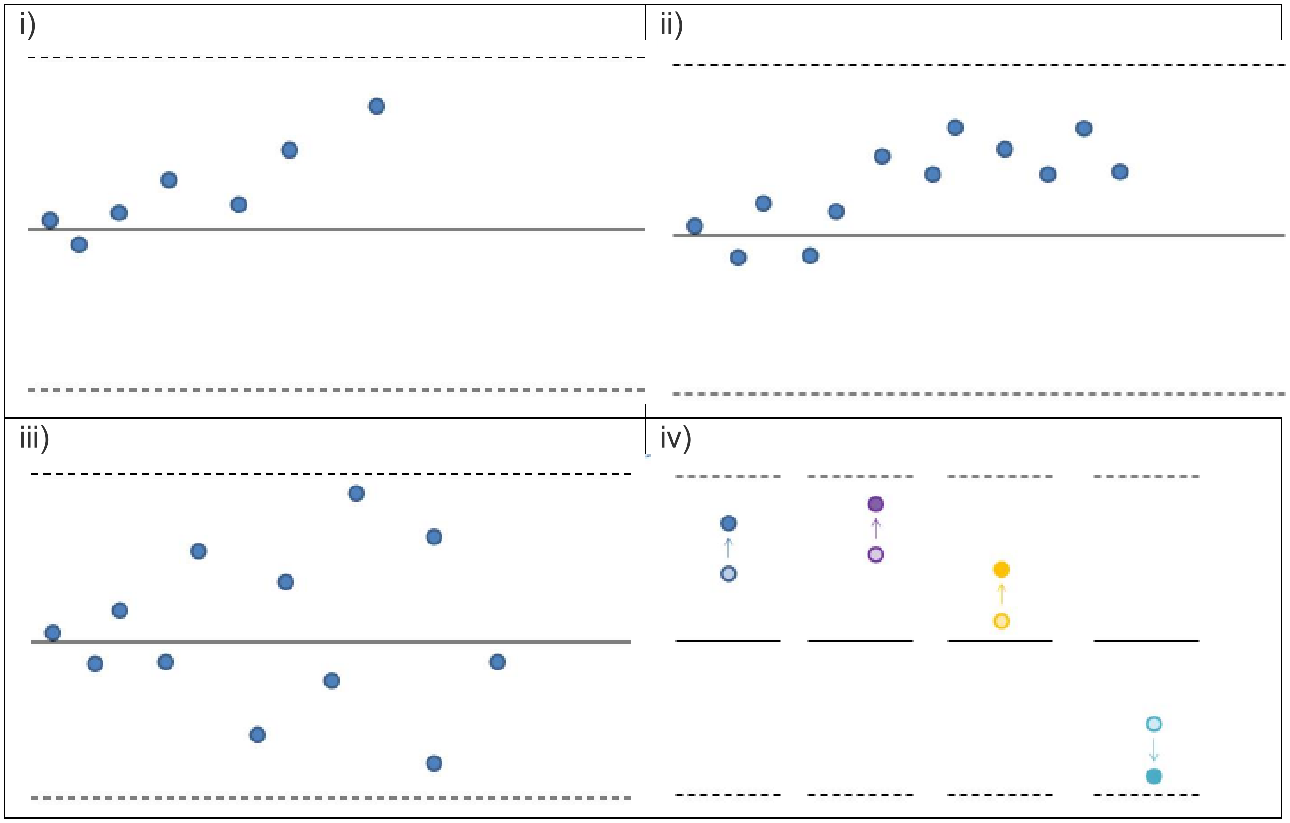


Figure 5. Unusual patterns that might indicate early warnings of a change even though the trigger (dotted line) has not yet been exceeded. i) a trend is observed that is approaching the dotted line. ii) a step change has occurred, the observations now appear to have stabilized around a new mean. iii) increasing variance (decreasing variance should also be investigated). iv) multiple indicators are all showing a shift towards undesirable condition.

4.3.4 Alternative Stable States

Ecosystems occasionally undergo rapid shifts as critical thresholds are reached and one stable state is replaced with another. For example, human induced eutrophication of lakes can occur abruptly when a critical nutrient threshold is exceeded (Scheffer et al. 2001). These sudden shifts are often difficult and costly to reverse and therefore triggers should be set much lower than the tipping point between alternative desirable and undesirable stable states. This subject was only briefly discussed during the workshop, but there are a number of papers that document the occurrence of these shifts as well as providing some insights into how to anticipate them (e.g., early warning signs and suitable indicators/triggers) (Scheffer et al. 2001; Briske et al. 2006; Dodds et al. 2010; Drake & Griffen 2010; Sheffer et al. 2012). In addition, after the workshop one participant prepared a useful summary of the concept of alternative stable states in the context of the workshop conversations (Appendix F).

4.3.5 More complex triggers

A number of more complex approaches were touched on briefly during the workshop. These should be explored further in subsequent workshops.

Variance based triggers: these are often used in manufacturing applications (Kenett and Zacks 1998) and have been successful at predicting catastrophic shifts among alternative stable states (Sheffer et al. 2009 & 2012).

Multiple indicators: Approaches can be extended to multiple indicators and applied in a multi-variate fashion but the math and the ability to communicate multi-variate results gets complicated quickly.

Reducing uncertainty with covariates: Using covariates to explain some of the variability (e.g., date to remove the seasonal noise) will help us to better evaluate triggers

Control charting approaches: Statistical process control methods are well developed in industry, particularly manufacturing and should be exploited for purposes of environmental monitoring. Many extensive references exist including: Kenett and Zacks (1998). There are a variety of control charts. There are charts for continuous or pass/fail data. Some plot the means and others plot the variance. More advanced charts weight more recent data more heavily than older data. In general, they consist of a chart with specified control limits on which data points are plotted as they are collected. The patterns observed are monitored for unusual behaviour with the goal of determining when the underlying process has changed, so that appropriate corrective action may be taken.

A number of different patterns may indicate non-randomness. For example for a simple $\bar{X}$-chart:

- A single point outside of +/- 3 standard deviations (an occurrence that would only be expected for ~ 1/100 observations)
- Two out of three points between +/- 2 standard deviations and +/- 3 standard deviations
- Six consecutive points increasing or decreasing
- A run of multiple consecutive points above (or below) the centerline.

The probability of observing each pattern if the process has not changed can be enumerated and appropriate triggers may be set depending on the risk tolerance of the program to Type I and Type II errors, respectively.

4.3.6 Power

The focus of this workshop was on setting triggers and what happens when a trigger is exceeded (e.g., confirmation of effect or management change). In a tiered program where the first response is to verify or confirm the effect, through additional more intensive monitoring, the risk of a Type I error (concluding there is an effect when in fact there is not) and subsequent costs (e.g., stopping development or increasing mitigation) is minimized by spending a smaller amount of money to confirm the result before taking more costly management actions. Several participants raised the inverse question, reminding the group of the risks of insufficient power (e.g., not detecting an effect when in fact an effect exists). They recommended more effort be put into robust design and power analyses, in essence they would like to see a parallel confirmation of 'no effect'.

4.4 How/when should values be incorporated?

Stakeholder values are important considerations in defining what we are ultimately trying to protect with environmental monitoring triggers. Therefore, it is important to ensure the approach to setting triggers is not disconnected with the values. Stakeholder values may or may not be directly measured or may be a culmination of multiple indicators. For example: Can we eat the fish? Can we drink the water? Can we swim in the lake? And will my grandchildren be able to? Understanding what we are trying to protect and identifying clearly unacceptable conditions is an important step. It is also important to ensure that the indicators and triggers being used do in fact link to value-based limits. It may be worth exploring whether or not triggers may be developed for stakeholder values directly. A big uncertainty identified at the workshop was at what stage stakeholder values should be considered in the process of setting triggers.

4.5 Accounting for complexity of ecological systems

As described in Section 4.4 the values we are trying to protect are often affected by multiple indicators at multiple scales. Understanding how to combine endpoints to provide a system level view of change to the values (i.e., the 'so what' interpretation) is a complex problem, but is a critical component to setting triggers. For example, we may use measurements of water quality as indicators, but the value we are trying to protect is having fish that are safe to eat. The linkages between the abiotic indicators and biotic response need to be better defined and verified if possible. This exercise is a necessary component of cumulative effects assessment. A number of questions were raised during this discussion and will require more research to provide sufficient guidance.

How do we roll up across multiple indicators at one site and one time? What about across multiple sites and multiple times?

- Vote counting methods are not a very good approach as they equally weight all indicators and don't account for the correlation among indicators.

How should triggers be applied across indicators, sites, and times? A problem in one indicator at one time is less of a concern than a problem in many indicators at multiple times.

- There are a number of analytical tools available to assist with these complex problems although they are generally under-utilized (e.g., confidence ellipses rather than intervals, multi-variate analyses, cluster analyses, redundancy analyses etc.)
- It is not necessary to pick a single method. It is possible to look at individual indicators independently as well as together using various multi-variate approaches.
- Some indicators may not result in an action but would inform decisions in a broader weight of evidence approach

4.6 Process/management frameworks

A regulator's perspective may be more focused on how triggers are contributing to the larger management framework and how project specific objectives tie into the regional context. Awareness of the different perspectives from project to regional should improve consistency and relevance of data collection and interpretation for all parties. While this wasn't the primary focus of this workshop, a presentation from the regulator's point of view on the North West Territory Response Framework for Aquatic Effects Monitoring resonated with the audience and helped to ground many of the more technical discussions on triggers. In particular the concept of clearly defined tiered action levels which was adopted by the participants (Section 4.3). This approach was also consistent with the tiered management response presented by Alberta Environment and Sustainable Resource Development with respect to their Surface Water Quality and Groundwater Management Frameworks.

5 Conclusions and Emerging Principles

5.1 Conclusions

Overall the workshop confirmed the need for further guidance in this area. While there were substantial differences in the current approaches, there were also many commonalities. The differences result primarily from the context, question(s) of interest, data available, and users (e.g., stakeholders). There was general consensus around using 95% threshold to define 'normal' although the details of the statistical approaches varied. Guidance cannot be overly prescriptive but several organizing frameworks emerged which should be useful in providing general principles to be considered under different scenarios. There is more work required to refine these organizing frameworks and flesh out some of the technical details. In addition there remains a lack of consistency in terminology used by workshop participants. A key insight from the workshop was the need for clearly articulated questions. Section 5.2 summarizes the emerging principles and preliminary considerations.

5.2 Emerging principles

At this point the guidance is not comprehensive, but a series of important questions to consider along with some emerging principles resulting from the workshop are provided.

5.2.1 Questions

The question and how it is asked is the primary driver for what approaches are taken when setting environmental monitoring triggers. More than one question may be assessed, but each should be clearly specified.

Are you comparing within one site, across local sites, or to regional sites?

Are you comparing to pristine, near pristine, or impacted sites?

Are you comparing particular time frames (e.g., before and after some impact)?

Are you comparing absolute values or relative change?

Is there a sequence to the questions, either by priority or the questions conditional on each other (e.g., if a, then evaluate b)?

5.2.2 Data availability

The availability of data (Table 5) will limit the ability to estimate normal and therefore answer all potential questions. If no suitable reference sites are available then we can only make comparisons within a site over time. If no pre-exposure data are available Before After Control Impact (BACI) studies cannot be completed, only comparisons of Control vs Impact (CI) or the site to itself over time.

Table 5. Data availability or lack thereof provides another potential organizing structure for guidance on developing environmental monitoring triggers.

	Historic data available?	
Reference data available?	*Y*,Y	*Y*,N
	N,Y	*N*,N

5.2.3 Defining Normal

The definition of normal should match the question. Therefore, there may be several definitions of normal, one for each question.

Normal defined by one or more reference sites:

The reference condition approach is very sensitive to how well you have matched your reference sites to your treatment site. It is very important to have rigorously defined comparable sites.

It is difficult to find a reference site that isn't impacted by something (e.g., hydro, logging, roads etc.). Minimally disturbed might be a more realistic comparison.

For comparison to other impacted sites, the same approach is used for selecting appropriately matched sites whether the reference site is a pristine site or an impacted site.

What if a reference site begins to drift, should it be removed from the reference population?

Normal defined temporally

Is pre-exposure data available? If so, is there sufficient data to adequately represent the baseline of interest?

If no pre-exposure data are available, the definition of normal should be updated annually until it stabilizes, or alternatively spatial replicates might be used in the early years.

Will the definition of normal be updated as more data are collected? Would a moving window be used, would all available data be used, would more recent data be weighted more heavily?

5.2.4 Triggers

There are two aspects to setting triggers: 1) what action levels or tiers are appropriate?; 2) for each action level or tier, what value of the endpoint or indicator should define the trigger?

Action levels or Tiers

Different studies will use a different break down of action levels or tiers and different terminology; however, in all cases they should be explicitly stated to improve clarity when developing triggers and communicating results. For example:

Don't worry be happy No action required
This is weird Low action level, this must be well within the 'Freak out levels'. Exceeding this level might trigger further data collection.
This is weird and abnormal Medium action level (could be a critical effect size), exceeding this level might result in a study to determine the extent and cause of the effect.
Freak out level. High action level. This corresponds to the level at which ecologically relevant changes (could be a critical effect size) or changes which impact stakeholder values (e.g., are the fish safe to eat?) occur.
PNR. Refers the point of no return, or very costly return.

Defining the trigger

Each action level for each question may have a unique trigger defined.

The trigger may be defined by a single exceedance of a set value or by an unusual pattern in an indicator (Figure 5).

Generally speaking, 2 standard deviations (or the 95% rule) represented the consensus for defining a "normal range".

Is the trigger applied to each observation, the mean, or the variance? The units used in the trigger must match the way in which normal is defined. For example, if the trigger is based on an annual mean, the definition of normal should also be based on what is normal for annual means, i.e., +/- 2 SD of the mean, or +/- $2 \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$.

It is possible to have triggers based on multiple indicators.

6 Recommended Next Steps

This workshop was the first in what could be several workshops and related research exercises to improve the guidance for how to develop environmental monitoring triggers. Preliminary guidance and key considerations are summarized in Section 5. This section describes the considerations requiring more discussion and recommends a series of actions to address some of the remaining uncertainties. Many of these items are the same as were identified at the end of the first day and discussed at length but not completely resolved on the second day of this workshop.

6.1 Parking lot

There were a number of further discussion items that warrant further attention but were not discussed specifically within this workshop:

1. How, or would, the approach change for questions focused on whether or not a project proponent is meeting Environmental Impact Assessment predictions?
2. Indicator selection, linkages to VECs, and roll up across indicators, space, and time. This subject was raised repeatedly and was only superficially addressed during this workshop as it was outside of the original workshop scope.
3. The specifics of defining triggers. Detailed guidance on the specifics of how to set triggers under different circumstances. More guidance on how to set interim (e.g., low/moderate action level, early warning) triggers to help prevent ever reaching the 'freak out levels' or critical effect size. What is the appropriate sequence for phasing triggers?
4. How and when to incorporate cultural values into trigger development and management structure? Consider using traditional or cultural indicators as triggers.

6.2 Documentation and calibration of the hierarchical pathways from indicators to VECs

Calibration of early warning indicators/triggers to value-based concerns (e.g., can I eat the fish) was identified as an important next step. Although some taxonomic groups are used as early warning indicators (e.g., algae and benthos), it is not clear how triggers based on these biota translate or scale relative to triggers using longer-lived biomonitors like fish. Suggestions for this calibration included overlaying and comparing dose-response curves for different biological groups and examining case studies of recovery to look for common trends and response patterns among different suites of biomonitors.

6.3 Retrospective analysis of existing data

Use existing datasets to evaluate:

- how many samples are needed for defining normal?
- what indicators are most relevant?
- what early triggers would be most effective?
- under what conditions each approach is most/least effective

RAMP data consists of 354 reach-years of data, with 5, 10 or 15 discrete 'samples' per reach. There is now a total of 3,526 samples. Historical datasets of impacted sites could be used to see if there is evidence of alternative stable states and if so if there were any early warning signs. Likewise, historical datasets showing recovery may be useful for identifying critical thresholds.

6.4 Next steps

There were a number of generalizations from this discussion

1. The type of trigger would depend on the question being asked and whether or not historical data was available
2. There are three tiers of questions
    a. What is normal for this site
    b. What is normal for the difference between this site and comparable reference site(s)
    c. What is normal regionally
3. Some measures are needed to determine when observations are approaching the limit of their "normal range". Statistical differences within a "normal" range should be interpreted with caution; consistent significant differences may be real but still within the "normal range"
4. There was general consensus that a measure corresponding to the number of SD (usually 2) could be used to define a "normal range". The standard deviation should stabilize within 8-10 sampling events (provided that there is not something else driving the variability), and that once you have sufficient data, triggers for a "normal range" could be set on a measure of variability
5. As a preliminary estimate, 2 SD of individual measurements will be a conservative estimate of 2 SD of means.

There are a number of issues that need to be explored further within the concept of triggers, including:

1. Once triggers are set, how often should they be adjusted, or should they?
2. How do you determine if "normal" varies over time and stabilizes around a new value?
3. How to compare indicators or triggers across levels of organization?

7 Literature Cited

Briske, D.D., S. D. Fuhlendorf, and F.E. Smeins. 2006. A unified framework for assessment and application of ecological thresholds. Rangeland Ecol. Manage. 59:225-236.

Dodds, W.K., W.H. Clements, K. Gido, R.H. Hilderbrand, and R.S. King. 2010. Thresholds, breakpoints, and nonlinearity in freshwaters as related to management. Journal of the North American Benthological Society. 29(3):988-997.

Drake, J.M. and B.D. Griffen. 2010. Early warning signals of extinction in deteriorating environments. Nature. 467:456-459.

Kenett, Ron S. and Shelemyahu Zacks. 1998. Modern Industrial Statistics: Design and Control of Quality and Reliability. Duxbury Press. Belmont, CA.

Kilgour, B.W., K.M. Somers, and D.E. Matthews. 1998. Using the normal range as a criterion for ecological significance in environmental monitoring and assessment. Ecoscience. 5(4):542-550.

Munkittrick, K.R., C.J.Arens, R.B.Lowell, and G.P. Kaminski. 2009. A review of potential methods of determining critical effect size for designing environmental monitoring programs. Environmental Toxicology and Chemistry. 28(7):1361-1371.

Scheffer, M., S. Carpenter, J.A. Foley, C. Folke and B. Walker. 2001. Catastrophic shifts in ecosystems. Nature. 413:591-596.

Scheffer, M., S.R.Carpenter, T.M. Lenton, J. Bascompte, W. Brock, V. Dakos, J.V.Koppel, I.A.Leemput, S.A. Levin, E.H. van Nes, M. Pascual, and J. Vandermeer. 2012. Anticipating critical transitions. Science. 338, 344-348.

Underwood, A.J. 1994. On beyond BACI: Sampling Designs that might reliably detect environmental disturbances. Ecological Applications. 4(1): 3-15.

Appendix A: Workshop Agenda

Trigger Development for Environmental Monitoring Programs

Workshop: June 11-12, 2014

Delta Lodge at Kananaskis
1 Centennial Drive
Kananaskis Village, Alberta
Conference call #: 888 350 3035 Conference room: 2426672
Goto meeting link: https://www3.gotomeeting.com/join/670312750 Access Code: 670-312-750

Workshop Objectives:

1. Develop a series of principles, and do's and do not's to be considered when triggers are developed.
2. We will do this through:
    a. Developing a common understanding of the various approaches currently in use to develop triggers and evaluate the strengths and weaknesses of these approaches
    b. Developing a common terminology around monitoring triggers

Day 1 Agenda:

Time	Task	Lead
8:00 (30 min)	coffee/snacks provided	
8:30 (30 min)	Introductions, workshop principles, agenda, and objectives	Lorne Greig / Darcy Pickard
9:00 (15 min)	Review of terminology	Kelly Munkittrick
9:15 (15 min)	Review of evolution of critical effect sizes and use of triggers in Environmental Effect Monitoring	Kelly Munkittrick and Fred Wrona
9:30 (45 min)	Presentation 1: UNB (Moose River)	Tim Arciszewski
	Presentation 2: Canadian Water Network (Saint John harbor)	Simon Courtenay
	Presentation 3: RAMP (Regional Aquatics Monitoring Program)	Bruce Kilgour
9:45 (30 min)	*Exercise 1: Complete worksheet – in subgroups*	Lorne/Darcy
10:15	*MORNING BREAK*	

Time	Task	Lead
(30 min)	*(snacks provided)*	
10:45 (60 min)	Presentation 4: Azimuth (metal mines)	Brian Pyper
	Presentation 5: Golder (diamond mines)	Tim Barrett
	Presentation 6: AESRD (Lower Athabasca Regional Plan - surface water framework)	Kim Westcott
	Presentation 7: Worley Parsons (Bow River)	Sally Hanearin
11:45 (30 min)	*Exercise 1: Complete worksheet– in subgroups*	Lorne/Darcy
12:15 (60 min)	*LUNCH (provided)*	
13:15 (60 min)	Presentation 8: Alberta Biodiversity Monitoring Institute	Jim Schieck
	Presentation 9: North Athabasca Oil Sands Area	Chantelle Leidl
	Presentation 10: Ontario Environment (Ring of Fire)	Keith Somers
	Presentation 11: NWT (triggers process)	Kathy Racher
14:15 (30 min)	*Exercise 1: Complete worksheet– in subgroups*	Lorne/Darcy
14:45 (30 min)	*AFTERNOON BREAK (snacks provided)*	
15:15 (30 min)	*Exercise 1 continued… summarize key insights from the day – in subgroups*	Lorne/Darcy
15:45 (30 min)	Sub-groups present key insights to the group – plenary discussion	Lorne/Darcy
16:15 (15 min)	Synthesis of the day, progress towards objectives, review of day 2 agenda	Lorne
16:30	ADJOURN	
evening	Group Dinner	

Day 2 Agenda:

Time	Task	Lead
8:00 (30 min)	coffee/snacks provided	
8:30 (15 min)	Review of agenda and workshop objectives	Lorne
8:45 (30 min)	Introduction to RAMP datasets (Water quality and benthic invertebrates)	Heather Keith / Bruce Kilgour
9:15 (75 min)	*Exercise 2: Data analysis exercise – in sub-groups*	Lorne/Darcy

Time	Task	Lead
10:30 (30 min)	*MORNING BREAK (snacks provided)*	
11:00 (60 min)	Data analysis - Plenary Discussion	Lorne/Darcy
12:00 (60 min)	*LUNCH (provided)*	
13:00 (90 min)	*Exercise 3: Drafting principles – plenary*	Lorne/Darcy
14:30 (30 min)	*AFTERNOON BREAK (snacks provided)*	
15:00 (45 min)	*Exercise 3 continued…*	Lorne/Darcy
15:45 (15 min)	Wrap up and next steps	Kelly / Lorne
16:00	ADJOURN	

Appendix B: Workshop Participants

Name	Affiliation
Allan Curry	University of New Brunswick
Arden Rosaasen	Areva
Bruce Kilgour	Kilgour Associates
Brian Pyper	Azimuth Consulting
Chantelle Leidl	AESRD
Darcy Pickard	ESSA Technologies Ltd.
Fred Wrona	Environment Canada
Heather Keith	Hatfield Group
Heather McMahon	University of New Brunswick
Janice Linehan	Suncor
Jim Schieck	ABMI
Kathy Racher	Wek'eeshii land and water board
Kelly Munkittrick	COSIA
Keith Somers	Government of Ontario
Kim Westcott	Environment Canada
Lorne Greig	ESSA Technologies Ltd.
Mark McMaster	Environment Canada
Meghan McEvoy	COSIA
Monique Dube	Shell
Murray Ball	AANDC
Neil Hutchinson	Hutchinson Environmental Services
Paul Jones	U Sask
Rainie Sharpe	Golder
Rick Lowell	Environment Canada
Rod Hazewinkel	AESRD
Sarah Depoe	AESRD
Simon Courtenay	Canadian Water Network
Tim Barrett	Golder
Tim Arciszewski	University of New Brunswick
Raju Neal Tanna	Worley Parsons

Appendix C: Day 1 Themes for further discussion

The following is a list of themes which workshop participants identified at the end of the first day. The suggestions were organized into a series of themes which were used to guide the day two discussions.

Categories of questions – how do approaches differ by question?
Project based vs. environment based monitoring programs > phased triggers.
Detecting change first is important even if the CES hasn't been reached

Normal -
What is normal? We have very short datasets compared to other disciplines
How do we define baseline conditions?
What data is used to define normal?
Relation of normal definition to value based triggers.
Reference sites – scale: regional, local

Methods / approaches -
How do we make decisions (Y/N) more objective? (stats)
Can we use available large datasets to help inform choice of triggers (e.g., use simulation to evaluate alternative triggers for a single dataset)
Defining your population – Tim B? may have missed his point here
Criteria for width of variation band - methods of calculation, mean, variance, how many SDs etc.).
Specifically what response measure is used and why? Mean across sites, within site, raw observations etc.
How triggers are defined (statistical: mean, CI, SD of mean, SD or CV directly; biological; regulatory)

Data availability –
What to do when lacking pre-development data?
What do you do when you don't have much data?
Revisit analyses of GSI data -
Compare/contrast the different approaches presented (Kelly's table should facilitate this)
Approaches were different but conclusions were not necessarily different, just framed differently – not a consensus
Would like to work through the example data in the group to reconcile the 5 different approaches/results we saw presented (Question clarification may help resolve these differences)

Power -

An exceedance triggers Confirmation of effect, however we don't have a parallel action for the opposite problem (i.e., not detecting an effect)
What might trigger more intensive monitoring (Exceeding trigger, Major process change by proponent)
How should we interpret lack of response? Would like to see more thought re. power/robust design.
Is adequate thought/effort going into design/power analyses

Values -
Identify clearly unacceptable conditions.
Identify incremental changes within value based limits.
Linking value based limits with incremental changes.
Relative importance of different endpoints.
Defining unacceptable, especially for biological data.
What are we trying to protect?
Once "normal" is exceeded you need to incorporate stakeholder values.
What are we protecting vs. what we are measuring?
Can we develop triggers for stakeholder values (e.g., can we eat the fish?)

Systems –
Disconnect between water quality and biological response.
Appropriate scale for change: appropriate indicator, what is normal, how to roll up, how actions are linked.
How do we roll up across indicators?
How to combine endpoints to provide a system level view of change.
Integration… or not
Integration is important for effective CEA (e.g., doing co-located/timed benthic and fish studies)

Process -
What is the process once you have triggers? (e.g. Kathy's presentation)
Project specific objectives – how do they tie into the regional context

Other -
What are the influences of interdisciplinary differences?
Would we design differently knowing that in 15 years we might not need the same design as we'll have enough data for other analyses (Marc-verify I captured his point)
After a project has been approved what is considered acceptable may be redefined putting proponent in difficult position.
Practical/Logistical constraints need to be weighed in determining approach

Appendix D: Day 2 Notes

This section consists of the detailed flipchart notes from discussions held on the second day of the workshop. They are organized into a series of sub-sections reflecting the key subjects discussed:

- The questions
- How is normal defined
- When do you re-define normal
- Different types of triggers
- Rolling up
- Next steps

The Questions

Change relative to prior conditions at same site
Change relative to conditions at reference site
Change relative to regional reference condition
t-test/anova-> is there a change that is big?
Pay attention to direction of a series of non-significant changes (control charting)
Need more than 1 reference site

- Is there a change
- Is there a big change
- Consistent with predictions
- Can't read last line

Has my site changed? (relative to itself) (2)

Has my site changed relative to local reference sites (2)?

Is there a change relative to regional reference? (2)

Is my site different from local reference sites? (1)

Where is my site relative to other impacted sites? (3)

[The order refers to the order of the questions. Originally we thought the order might differ depending on whether historical data were available or not but in the end the group gave us the same order regardless. I'm not sure I agree, it seems like there was a natural hierarchy which this ignores: site, local ref, regional ref). Clarity is required to help communicate the difference between: different and change. Basically if looking at change relative to reference sites we are looking at whether they are tracking each other, whereas if we are looking at 'difference' relative to reference sites we are looking at whether they are the same – do they have the same absolute numbers as the reference sites.]

Brian's figure:

	mean	change
Site	x	BA
Local reference	CI	BACI
Regional reference	MVE	MBACI

Simon's figure:

	Historic data available?	
Reference data available	Y,Y	Y,N
	N,Y	N,N

[What should you do in each of these scenarios? This is the framework Simon suggested for laying out the principles.]

How is Normal Defined?

If reference sites are available, these are used to define normal

Normal must be defined in the same units (one observation vs mean observation)

Temporal aspect to normal

Variation in year to year means
Normal = baseline (pre-exposure)
Normal = upstream of impacts
Which years should be included?

Unimpaired vs least impaired, 'pristine' or 'natural' may not exist or be realistic. 'minimally disturbed' might be a more realistic term.

When do you collect the data? Ideally you would collect regional reference data from the start

Approach to each question depends on: available data (pre, reference)

In what situations are the questions conditional (e.g., local ref first, then regional)

Cumulative [what does it really mean?]

- Multiple indicators

- Multiple time
- Multiple space

Regional reference sites incorporate cumulative effects

Reference for oil sands might still be impacted by other development (hydro etc.) hard to get a control that isn't impacted by something

If no historic data

Default CES
2SD
Can still do spatial comparisons, especially important in early years
Use data from the literature
Expect the variance to stabilize in 8-16 years same idea spatially, Assuming distribution isn't changing
The questions!

What should you do in the 1st year in absence of data

~8 ref sites

If random sample from population (i.e., appropriately selected reference sites)

With small sample sizes should we use t distribution instead for SD multipliers

Don't necessary have to worry so much about the stats as other sources for trigger
Put a CI on the trigger (e.g., 2 SD +/-CI)

How many SD's?

2 is common
3 or 4.5 was used in some of the examples
If small n and CI on 2 SD may be similar

Choice of X SD depends on tolerance of your program
2 (or 1.96) comes from the 95% rule, this may be a better number to use for communication

Figure illustrating tolerance limits on trigger

When do you redefine normal

Continuously while you think you are still 'normal'

Would not expect to see many years in a row, above or below the mean

If variance is not stabilizing something may be going on (ref BC papers)

What if a reference site starts to drift? Do you drop it?

Have to be careful not to be too loosey goosey with rules on changing normal. If you don't like what you see.. you change it—is that ok?

Normal is not fixed in reality

Environment Canada uses moving 30 year window as 'normal'

Depends on the question, often a fixed benchmark is used

May have different benchmarks for different questions, must clearly articulate the questions

Could use a rolling average

Use 'clean' pre-development data (don't see why you would update -assuming it is available)

Can do both

- 'near pristine' benchmark
- Incremental change relative reference sites

With no pre-data

- Update annually until it stabilizes
- May want to utilize spatial data (replicates) to supplement in the early years

Different types of triggers

Trigger action levels

- Early warning
- Low action
- FOL [freak out level]
- Etc.

Detecting a significant change does not necessarily mean you are near the FOL

Looking for trends to see if approaching FOL

Triggers can be lined up with the questions (1-is my site different etc.)

Indicator selection problem, do some respond more quickly

How close to predicted are you?

Low action trigger may be at lower level than CES. IF collecting more data is the 'action' it will inform the next step.

Low level trigger must be well inside the FOL
It may be that an ecologically relevant FOL is not reached when sub-indicators are exceeded

Some indicators may not result in an action but would inform decisions in a broader weight of evidence approach

Multiple indicators trending in the same direction at the same time

Need clear link between indicator and VEC

Indicators should be responsive

When there is a tipping point between 2 stable states, need to set trigger well short of tipping point. Variance based triggers may be most useful here

Could RAMP data be used to help understand how many samples needed for defining normal (RAMP~450 samples divided by 10 to get annual means)

Figure of alternative early triggers

If we can understand the site level noise better we can better evaluate triggers (compare observed to predicted)

If trend exists is it explained by covariates? If no, is it outside normal.

Using covariates to explain some of the variability (e.g., date) –recall Kelly's example of focusing on Aug 1 to remove the seasonal noise

Rolling up

Necessary for cumulative effects assessment

Roll up for multiple indicators (at one site, one time), multiple sites, multiple times

A problem in 1 indicator at 1 time is less of a concern than a problem in many indicators at multiple times

Roll up – the 'so what' interpretation

2 parts:

1) Multiple indicators showing response
2) Space –showing response across multiple sites

Multiple indicators in different categories is very concerning

Vote counting method doesn't account for correlation among indicators

Lots of [analytical] approaches out there, just not common practice yet

- Deviations from the mean, on multiple indicators @ same time (can be measured e.g., 95% ellipse)
- Not one [approach] or the other, need to look at indicators one at a time as well as together
- Don't have to pick just one method can look at it 1 at a time and together
  - 1 indicator ANOVA
  - 2 indicators MANOVA
  - Multi-variate analyses, try dropping one variable at a time and see what happens
- Cluster analyses may help ID redundancies in indicators
- Redundancy analyses
  - Multiple y's and x's
  - Could use a step change as a predictor variable
  - Allows elaborate hypothesis testing

What is the connection between 'lower' and 'higher' indicators?

If 'lower' indicators start to be triggered… would expect to see 'higher' indicators (e.g., fish) start to show change (lower and higher refers to the hierarchy from raw indicators to VECs)

Next steps

Reference sites

Spatial/temporal variance in reference areas in determining change vs. difference

Questions

Regulatory context may need to be considered
Project proponent – interested in whether you are meeting predictions
EIA predictions

Indicators/roll up

Triggers for multiple components, how do we actually do?
How to roll up individual sits to broader region (E.g., Athabasca)
Roll up
Everything starts with your indicator → want to see the hierarchical pathways laid out among VECs
Calibration of early warning indicators/triggers to higher level [value based concerns – can I eat the fish] need to confirm the links
Pathways from indicators to VECs
What should I do at regional scale (tribs, watersheds, regions)
What assumptions are involved in regional analyses
Indicators
    Ecological relevance & time
    Adverse outcome pathways
Action levels tied to questions

Triggers, early warnings & the details
Mechanics of it (details)
More specifics are needed next
Need to focus more on interim triggers (avoid getting close to FOL)
How do we set the interim triggers (25%, 75%)
How the triggers are phased, is there an appropriate sequence
Early warnings
Phased management action with multiple indicators
How far can you go before you go to the next stable state and can you come back [at what cost]

Key Considerations and guidance
Normal range – decent consensus [somewhat of a surprise, that there was as much agreement as there was]
Can't be overly prescriptive
Clearly articulate questions
Guidance on how we select reference sites
Language is still a problem even in this room
How similar / different are the approaches
    Lots of similarities
The options depend on your specific problem. There are commonalities but won't fit all.
    Data available
    Stakeholder involvement
    Questions (what they are)

Values
Consider traditional/cultural indicators for use as triggers

When do we bring in the value conversation
How do you incorporate cultural values into triggers and how to incorporate into management structure
Are there reasonable size steps we can agree on and we can communicate?
Don't worry be happy
This is weird → low action level
This is weird and abnormal → medium action level
Freak out level
SOL

Other

How would you take EEM data within site and re-evaluate (based on conv.) [not sure what this point was]
Would I design any different? Still do upstream/downstream etc.?
How do we build this process?

Actions

Retrospective analyses
Patterns among mines etc. to see if we see evidence of alternative stable states
What is the most effective tool based on existing datasets
Could also use datasets showing recovery
Do retrospective studies to test triggers

- What indicators are most relevant
- What early triggers
- Would like to try testing approaches with data (try breaking it)

Historical data

- We have short data sets try to use any available data even if from other studies, don't want to waste it

Appendix E: Brian Pyper Post Workshop Summary

The following is a set of post workshop notes provided by Brian Pyper regarding the proposed organizing structure in Table 4. Further discussion with the group is required to synthesize these notes into guiding principles regarding the approach to setting environmental monitoring triggers.

1. Obviously, we are concerned with triggers, and this table says nothing about triggers per se. Yet it is exactly the type of structure (IMO) required to guide discussions of triggers by defining questions of interest in the context of available data.
2. For each box, we could easily specify (write out in symbols) a statistical model for the data at hand and the key parameter of interest. For clarity, I would specify the complete (hypothesized) structure for the case where we have multiple exposure sites, multiple reference sites, and multiple years of data (i.e., the ideal data for box 6). This would denote the “global model” from which all other boxes are shown to be a generalization (simplification) thereof.
3. In the ideal, this “global model” would be a mixed-effects model that:
    - (a) accounted for the hierarchal structure of site/year units with likely unbalanced subsampling therein (e.g., N=20 replicate fish sampled for GSI at one site/year and N=10 fish the next year);
    - (b) modeled via a parametric distribution the differences in means (and possibly variances) among sites in the context of the “meta-population” of interest;
    - (c) incorporated explicit correlation structures for space and time (i.e., sites and/or years are not “independent” given spatial/temporal processes unrelated to impacts); and
    - (d) incorporated terms to test explicit hypotheses (e.g., exposure site A has recently changed; exposure site A has recently changed relative to the local reference; exposure site A is different from regional references, etc.)
4. Again, this is only the ideal, meant to provide a framework and context for understanding specific applications. (Analytically, I’ve worked extensively with such conceptual models and in the most complex case, with spatial/temporal correlation structure, the only tool I’m aware of is Bayesian hierarchical modeling in WinBugs, for example).
5. More practically, it serves as an analytical framework for acknowledging what is missing in most conventional analyses due to limited years and/or sites: potential temporal/spatial autocorrelation in the data (ignoring which could elevate Type I errors).
6. For example, I included “intervention analysis” in the table because most time-series statisticians would think of the “change” question in this context; however, the reality is that most monitoring datasets are too short to support reliable estimation of time-

series components (e.g., ARMA terms), so we generally omit them (either knowingly or ignorantly).

7. Note that incorporating deterministic temporal structures such a month of sampling (Tim B's application, something Kelly emphasized) is the simple addition of a covariate to any box.
8. Triggers: I think the utility of the above table was best exemplified by the divergent conclusions of the trigger examples we applied. At one point, it was suggested that each approach resulted in a more or less similar conclusion, to which there was large consensus, but that was hardly the case. The approaches/conclusions differed dramatically.
9. The discussions focused primarily on "statistical" triggers rather than "effect size" triggers. Either way, the above table applies.
10. In short, the predominant statistical trigger was considered as 2*SD, but SD depended on the question of interest. Here, it's useful to think of the four key "components of variations" (i.e., sources of natural variation or "noise") for the "global model": (1) there is variation in means among sites across time (call this SD[sites]); (2) there is variation in means among years across sites (call this SD[years]); (3) there is variation in means that is specific to each site/year combination (call this SD[site*year]); and (4) there is variation within each site/year combination due to sample replicates or "subsamples" (call this SD[samples]).
11. To summarize, as best as I can recollect:
    - Tim A's approach was concerned with temporal change in GSI at the exposure site (Box 2), so SD = the standard deviation of previous annual means at that site (i.e., SD[years] for the exposure site). Data from the reference site were used "qualitatively" to assess changes in the context of a possible local trend. Another approach mirrored this: something like 4*SD was applied to individual fish measures (SD[samples], rather than annual means), yielding a similar conclusion that years 14-16 were problematic (change detected).
    - Tim B's approach was concerned with annual differences in mean GSI between the exposure and reference sites (Box 3, but annually), so SD = SD[samples]/sqrt(N). The approaches applied by Bruce Kilgour and Keith Somers were similar. The opposite conclusion was reached: GSI was different until years 14-16.
    - Brian P's approach was concerned with temporal change in GSI at the exposure site relative to the reference site (Box 4), so the analogous SD was the standard deviation of previous differences in annual means (SD[site*year]). A different conclusion was reached: change was evident by year 11 (conclusive for years 14-16).
    - The second-morning discussion of fish harvest in lakes (Bruce) was basically Box 5. Here, SD of interest = SD(sites).
    - In summary, the methods discussed covered Boxes 2-5, but I think that Boxes 5 and 6 are what folks will ultimately be interested in.

Appendix F: Keith Somers Post Workshop Summary

The following summary was prepared by Keith Somers following the workshop for the purpose of briefing others on the outcome of the meeting.

Environmental Triggers Workshop

A workshop on "Trigger Development for Environmental Monitoring Programs" was organized by Canada's Oil Sands Innovation Alliance (COSIA) to investigate the current use of environmental triggers in monitoring programs by:

- identifying the types of environmental triggers currently in use;
- understanding how these triggers were developed; and,
- understanding how environmental triggers are being utilized.

To achieve these objectives, COSIA invited a group of scientists with experience in environmental monitoring to a two-day workshop (June 11-12, 2014) in Kananaskis, AB.

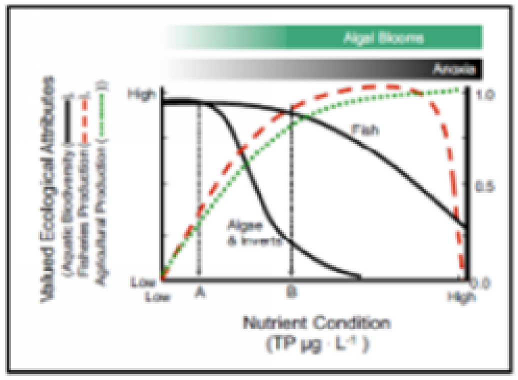


Figure from Stevenson (2014 Ecological assessments with algae: a review and synthesis. J. Phycol. 50: 437-461) showing that the biodiversity of various biological groups changes at different points along a gradient in nutrient concentration. A nutrient criterion at point A would protect all biodiversity whereas a criterion located at B would protect fish biodiversity, but not the biodiversity of algae and invertebrates.

Environmental Triggers Workshop

The two-day workshop began with a series of invited presentations from individuals who were involved with environmental monitoring programs that utilized triggers and thresholds:

1 – Moose River, Ontario – Tim Arciszewski
2 – St John Harbour, NB – Simon Courtney
3 – RAMP – Alberta Oil Sands – Bruce Kilgour
4 – Metal Mines – Brian Pyper
5 – Diamond Mines – Tim Barrett
6 – Lower Athabasca Regional Plan – Kim Wescott
7 – Bow River – Sally Hanearin
8 – Alberta Biodiversity Institute – Jim Schieck
9 – N Athabasca Oil Sands – Chantelle Leidl
10 – Ontario Biomonitoring – Keith Somers
11 – NWT Mining – Kathy Rachar

View of the court yard and mountains at the Delta Lodge, Kananaskis, AB.

Environmental Triggers Workshop

The presentations revealed that different groups were using similar terminology, but they were asking different questions. Subsequent discussion identified the following 3 critical questions using 'lake' as the system of interest:

1. Has my lake changed?
2. Have similar lakes near my lake changed in a similar manner?
3. Has there been a regional change in lakes that follows the same pattern?

View of the mountains from one of the walking trails at the Delta Lodge, Kananaskis, AB.

The questions being asked were influenced by how each question was posed, in what order, the experimental design (e.g., before-after, upstream-downstream, reference condition approach, etc.), the statistical tests, and how uninteresting variation (or noise) was removed. The discussion led to general agreement that the questions needed to be clear, there was value in ordering the questions, and the concept of the normal range of variation among reference sites was a useful context for evaluating the importance of an observed change.

Environmental Triggers Workshop

The late afternoon on the first day focused on different approaches used by each of the participants to set triggers or thresholds for 3 data sets provided by the workshop organizers. The results of this exercise underscored a general inconsistency in how triggers and thresholds were applied – with the exception that all participants were comfortable with the concept that 2 Standard Deviations (SDs) was a reasonable trigger if no limits of acceptable change were available before a study was initiated (see subsequent slides for details on this concept).

Another view of the court yard and mountains at the Delta Lodge, Kananaskis, AB.

The workshop facilitating team also asked each of the participants to list the critical questions that they hoped would be answered by the workshop. The list included questions relating to: best experimental designs, using multiple triggers simultaneously, early warning indicators, nonparametric tests, all indicators are not equal, assessing impacts objectively, effect sizes, assigning probabilities in an assessment, combining results from different indicators – most of the questions focused on data analysis and interpretation.

Environmental Triggers Workshop

At the start of the second day the original agenda was discarded and the group collectively decided to examine the most common, critical questions from the first day and suggest solutions based on open discussions, as well as through break-out groups.

The methods described by Kilgour et al. (1998) were used to illustrate the concept of the normal range (i.e., 2 SDs) and demonstrate statistical tests that objectively assign probabilities for unclassified locations (subsequent slides summarize the basic concepts underlying these group discussions).

Break-out group discussion at the Environmental Triggers Workshop in Kananaskis, AB.

The workshop continued late into the second day with the general agreement that a second workshop should be held to address the issue of calibrating summary metrics for the biological groups that are routinely used in bioassessments. Although some taxonomic groups are used as early warning indicators (e.g., algae and benthos), it is not clear how triggers based on these biota translate or scale relative to triggers using longer-lived biomonitors like fish. Suggestions for this calibration included overlaying and comparing dose-response curves for different biological groups and examining case studies of recovery to look for common trends and response patterns among different suites of biomonitors.

Environmental Triggers Workshop

To assist in understanding how environmental triggers are used, there was a discussion of the theoretical underpinnings of evaluating ecosystem condition, its common states, and how ecosystems change from one state to another.

The figure on the right illustrates how an ecosystem can change from one state (the black ball) to an alternate state (the grey ball) by moving from one normal operating range (the left cup) to a new operating range (the right cup).

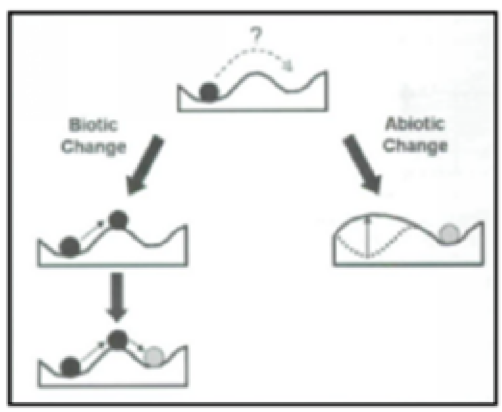


Figure from Briske et al. (2006) A unified framework for assessment and application of ecological thresholds. Rangeland Ecol. Manag. 59: 225-236.

The figure presents 2 scenarios showing that a biological change can move an ecosystem to an alternate state, but that ecosystem can return to its previous state by reversing the biological change (e.g., stocking a fish species that was removed by over-harvest). By contrast, an abiotic change can move the ecosystem to a new state, but return may not possible if the original condition has changed.

Environmental Triggers Workshop

Using this simple ball-and-cup example, a ball remains in a cup because negative feedback (NFB) effects are greater than positive feedback (PFB) effects. A 'trigger' is an event or action that causes the ball to move to an alternate state (i.e., when PFB = or > NFB). A 'threshold' is a region of rapid change traversed by the ball as it moves to a new state. Intervention is generally required to move the ball back to its original state, although this is not always possible.

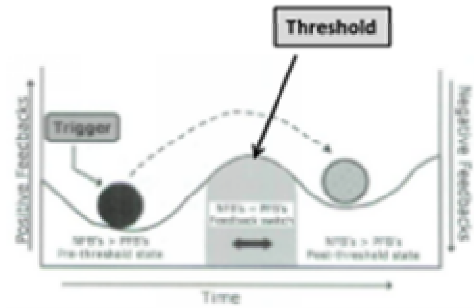


Figure from Briske et al. (2006) A unified framework for assessment and application of ecological thresholds. Rangeland Ecol. Manag. 59: 225-236.

Lake acidification is an example of this model. The black ball on the left represents a lake in its normal state. As the pH of that lake changes, other characteristics of that lake also change. At some point in time, the lake transforms to exhibit the characteristics of an acidified lake (the grey ball) after passing through a period of rapid change (i.e., the threshold). Intervention (e.g., by liming) could return the lake to its original chemical state. A pH of 6 has been proposed as the threshold that results in the rapid loss of species and associated biodiversity. A downward trend in the pH of an unimpacted lake could be a trigger indicating that a change in lake condition will occur when the pH falls below 6.0.

Environmental Triggers Workshop

The figure on the right illustrates that continued perturbation can sequentially move an ecosystem to exhibit greater impairment once a threshold has been crossed – from structural changes, to species losses, to changes in ecosystem function, and ultimately to the extinction of all species.

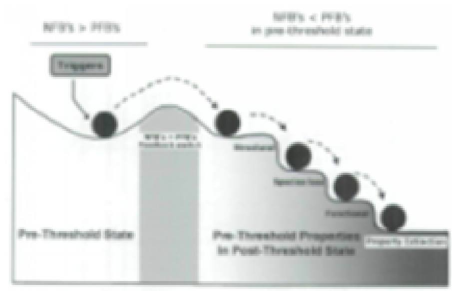
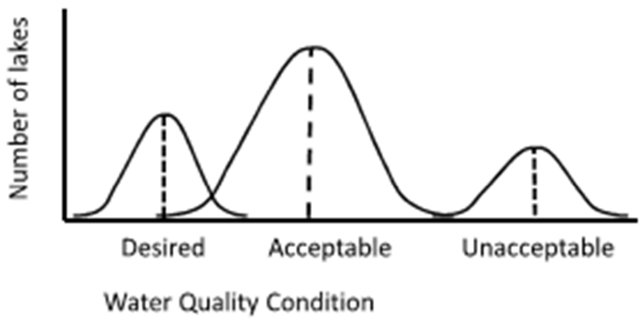


Figure from Briske et al. (2006) A unified framework for assessment and application of ecological thresholds. Rangeland Ecol. Manag. 59: 225-236.

The experimental acidification of Lake 223 at the Experimental Lakes Area (ELA) illustrates this concept (from Schindler et al. 1985).

Year	pH	Observed change
1976	6.5	alkalinity decreased; otherwise within normal range
1977	6.1	most results within normal range
1978	5.9	Mysis numbers fall; fathead minnows did not reproduce
1979	5.6	filamentous algae appears; fathead minnow numbers down
1980	5.6	some copepods and cladocerans rare; crayfish reproductive failure
1981	5.0	several zooplankton species disappear; white sucker reproduction failed
1982	5.1	lake trout condition poor; crayfish numbers down
1983	5.1	lake trout condition worse; crayfish, leeches and mayflies gone

Environmental Triggers Workshop

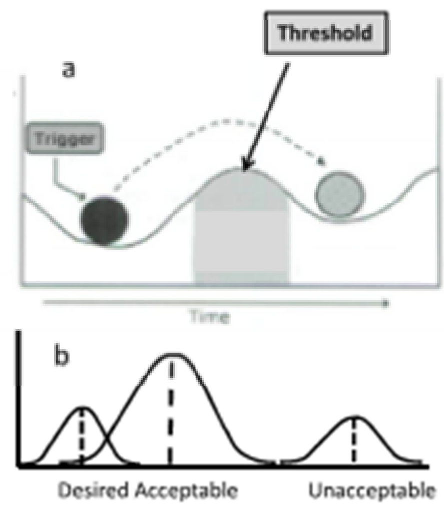


This figure provides a different perspective on the idea of alternate stable states, triggers and thresholds. Here, each curve represents the distribution (or histogram) of a series of lakes based on water quality measurements (say, nutrient concentration). The left-most curve represents a histogram of lakes exhibiting excellent water quality associated with pristine conditions – what we might call the "desired" water quality condition.

In the middle of the figure is a histogram of lakes that have experienced water quality changes from the desired condition, but these changes are "acceptable". On the right is a third distribution representing lakes that have experienced a large change in water quality and these lakes are in an "unacceptable" state.

Environmental Triggers Workshop

If we display the ball-and-cup figure (a) and the histogram figure (b) together, the desired and acceptable conditions in (b) represent the normal operating range for the black ball in (a). The unacceptable condition (shown in b) is illustrated by the normal operating range for the grey ball in (a). This example shows that the unacceptable condition is achieved by moving the black ball from the desired/acceptable condition across a threshold to the location of the grey ball representing an unacceptable alternate stable state.



Moving an ecosystem from an unacceptable state to a desired or acceptable state generally requires intervention and a significant investment of resources over time. For example, legislated reductions in atmospheric emissions of sulphur dioxide have contributed to the recovery of many of Ontario's historically acidified lakes. Recent concerns about decreasing calcium concentrations suggest that the full recovery of acidified lakes make require additional intervention.

Environmental Triggers Workshop

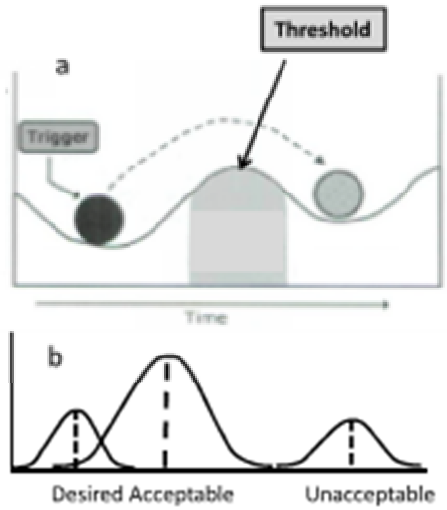


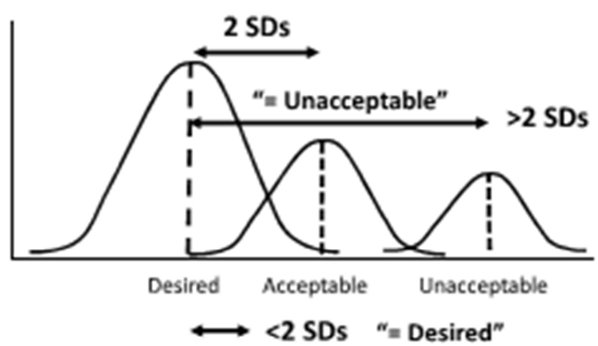
Figure (a) illustrates that a 'trigger' is a change in the normal operating condition that provides an early warning of a pending shift from a desired or acceptable state across a 'threshold' to an unacceptable state.

The vertical lines associated with each distribution in (b) indicate the median or average condition for each state: desired, acceptable and unacceptable.

The difference between the average values for the desired and acceptable conditions indicate the limits of acceptable change. Generally Guidelines and Objectives are used to indicate the limits of acceptable change.

Environmental Triggers Workshop

In the absence of Guidelines or Objectives, an alternate approach to characterize the normal operating range (also described as the typical range) involves sampling a large number of minimally impacted reference sites. This 'Reference Condition Approach' (RCA) estimates the distribution of sites in the 'Desired' condition (i.e., the left-most histogram).



Kilgour et al. (1998) proposed that the distribution of reference sites in the 'desired' condition can be used to estimate the limits of acceptable change. Since 95% of these reference sites will fall within the region defined by the mean plus-or-minus two standard deviations (SD), Kilgour et al. suggested that 2 SDs should be used to define the limit of acceptable change. They also described appropriate statistical tests to evaluate whether an unclassified site is significantly greater than 2 SDs, and hence, belongs to the distribution of sites in the 'unacceptable' condition; or, is significantly less than 2 SDs, and hence, is equivalent to the reference sites in the 'desired' condition.

Environmental Triggers Workshop

References:

Briske, D.D., S.D. Fuhlendorf and F.E. Smeins. 2006. A unified framework for the assessment and application of ecological thresholds. Rangeland Ecol. Manage. 59: 225-236.

Kilgour, B.W., K.M. Somers and D.E. Matthews. 1998. Using the normal range as a criterion for ecological significance in environmental monitoring and assessment. Ecoscience 5: 542-550.

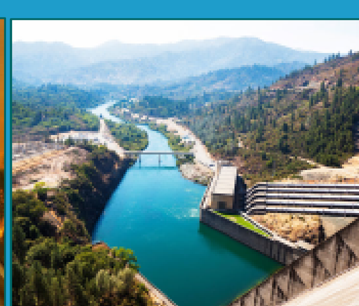
Schindler, D.W., K.H. Mills, D.F. Malley, D.L. Findlay, J.A. Shearer, I.J. Davies, M.A. Turner, G.A. Linsey and D.R. Cruikshank. 1985. Long-term ecosystem stress: the effects of years of experimental acidification on a small lake. Science 228: 1395-1401.

Stevenson, R.J. 2014. Ecological assessments with algae: a review and synthesis. J. Phycol. 50: 437-461.

Final Report

Environmental & Cumulative Effects Assessment

Climate Change Adaptation & Risk Reduction



Aquatic Species at Risk & Water Resource Management

Terrestrial Ecology & Forest Resource Management