diversity

MDPI

Review

Principles and Challenges for Multi-Stakeholder Development of Focused, Tiered, and Triggered, Adaptive Monitoring Programs for Aquatic Environments

Kelly R. Munkittrick [1,*], Tim J. Arciszewski [2] and Michelle A. Gray [3]

[1] Department of Biology, Wilfrid Laurier University, Waterloo, ON N2L 3C5, Canada
[2] Environmental Monitoring and Science Division, Alberta Environment and Parks, Calgary, AB T2L 2K8, Canada
[3] Canadian Rivers Institute and Faculty of Forestry and Environmental Management, University of New Brunswick, Fredericton, NB E3B 5A3, Canada
* Correspondence: kmunkittrick@wlu.ca

Received: 29 June 2019; Accepted: 1 September 2019; Published: 4 September 2019

Abstract: In Canada, there is almost 30 years of experience in developing tiered and triggered adaptive monitoring programs focused on looking at whether environmental concerns remain when pulp and paper mills, or metal mines, are in compliance with their discharge limits. These environmental effects monitoring programs were based on nationally standardized designs. Many of the programs have been developed through multi-stakeholder working groups, and the evolution of the program faced repeated frictions and differing opinions on how to design environmental monitoring programs. This paper describes key guidance to work through the initial steps in program design, and includes scientific advice based on lessons learned from the development of the Canadian aquatic environmental effects monitoring program.

Keywords: environmental effects monitoring; adaptive management; monitoring study design

1. Introduction

Human activities affect the environment. Greater recognition of project-specific and cumulative influences has resulted in the proliferation of monitoring programs to address multiple purposes. A common form of environmental monitoring is examining exposure areas for residual effects when proponents and operators of activities comply with permitting requirements. In addition to these effect monitoring programs, others can also be designed, including regional ambient programs. In Canada, industry-funded and government-approved environmental effects monitoring (EEM) programs developed for pulp and paper in the early 1990s [1] and metal mining in the late 1990s [2]. The Canadian metal mining EEM program [3] was developed through a multi-stakeholder committee process which took a number of years. The issues creating tension and conflict occurred during discussions involved in adjusting the monitoring approach in multi-stakeholder discussions in Chile [4] and Brazil [5], as well as the evolution of the regional oil sands monitoring program in Canada [6]. Examples of points of conflict include whether the regulated industry can sit at the table as part of the discussions surrounding development, and whether the monitoring should focus on stressors, effects, or values. From a philosophical viewpoint, there are dozens of additional issues to consider as any new regional monitoring program is developed that impact the design and interpretation [7]. Each of the issues requires extensive discussion to resolve an approach, especially when various stakeholders are involved in the development. Such discussions take time to develop trust among participants, and to

develop relationships that enable transparency, inclusiveness, clarity, and a common understanding of the objectives, limitations, and deficiencies of approaches, the data generated, and the decisions which can be supported by the monitoring program.

Successful monitoring programs require clear and common understanding of how funding is allocated and the goals of the program. Discussions on each of the issue require time to resolve, since each question has multiple answers, and different stakeholders have different perspectives; based on our experience, these perspectives can be expressed as multiple alternative answers to a series of questions (Table 1). There will be conflicting opinions for each issue that will affect the design of the program. Resolving the misalignments can be challenging, and takes time, patience, and often, good facilitation.

Table 1. Issues and concerns for developing a multi-stakeholder aquatic monitoring program.

Program structure and administration	1. The program should be developed by	Industry, government, multi-stakeholder
	2. Industry should be allowed input or review	Yes, No
	3. Studies should be paid by	Industry, government, other
	4. Requirements should be	Written into a permit, a regulation, OR are voluntary
	5. Goal is to	Identify impaired areas inform adaptive management, prevent problems, restore conditions
	6. Management decisions required	Within 10 years, within 5 years, soon
	7. Monitoring program focus	Site-specific, regional, national
	8. Participants measure the same things	Yes, No
	9. Effort varies depending on the level of concern	Yes, No
	10. Is there a core program, with site-specific requirements	Yes, No
	11. Will results be used to change intensity or requirements	Yes, No
	12. Study designs approved by	Government, peer review, industry
	13. Results received by	Program office, regulator, industry
	14. If the study is inadequate	Program resample, proceed to next cycle
	15. If a change is detected	Regulatory action, management action, change intensity of monitoring
	16. Decision will be based on	Ecological integrity and biodiversity, impact on water use, fish use or the fishery, increasing change
Program Objectives	17. Focus of the program is	Human perspective and use, ecological perspective, both
	18. Main purpose	Monitoring, surveillance, assessment, prediction
	19. Main focus	Unknown stressors, cumulative stressors, specific development
	20. Design to look for	Stressors, effects, protecting values
	21. Is research allowed	Yes, No
	22. Program looks for	Change, effects, impacts
Study design	23. Pre-development data	Yes, No
	24.If sites that are confounded or dangerous to sample	Alternative methods, monitoring is not necessary at all sites
	25. When change is detected	Trigger studies, identify cause, fix change
	26. Consequence of seeing a change	Collect more information, define cause, management decision
	27. Decisions will be based on	Individual results, a pattern of responses, weight of evidence
	28. Reference data	Historical data, temporal within site, comparable reference sites, gradient
	29. Regional reference data	Within program, external
	30. Natural variability	Not relevant, detected, understood
	31. Confounding factors should be	Avoided, detected, identified
	32. Focus	Species used by people, most exposed, all are important
	33. Level of confidence desired (alpha)	1 in 10 (0.10), 1 in 20 (0.05), OR 1 in 100 (0.01)
	34. Level of power desired	80%, 90%, 95%
	35. How big a change do you want to detect	Statistical difference, predefined effect size, two standard deviations

Table 1. *Cont.*

Data Analysis and reporting	36. Site data will be analyzed by	Industry, government, an independent agency, the public
	37. Broader program data will be analyzed by	Industry, government, an independent agency, the public
	38. Data analyses need to be available within	1 year, 2 years, 5 years
	39. Study results will be	Public, only summaries are public, limited circulation, confidential
	40. Summaries will be available	Internet, public presentations, public reports, scientific literature
	41. Data summaries will include	Public summaries, evaluations against triggers, comparisons against predictions
Interpretation	42.Will any change be interpreted as significant	Yes, No
	43. It is important to decide what an effect	Is, is not
	44. The importance of a change will be decided by	Pre-defined triggers, best professional judgement, scientific or peer review
	45. Does unacceptable change have to be decided in advance	Yes, No
	46.The acceptability of a change is decided by	Government, stakeholders, it is a negotiation
	47. An unacceptable change should	Be fixed, monitored more closely, it depends
	48. An ecologically relevant change is	Any change, a change that is getting worse, a change that threatens biodiversity
	49. An unsustainable change is	Any change, change that threatens growth, reproduction, survival and use, any change is important
	50.The monitoring system should be adapted	When needed OR every 3, 5 OR 10 years

Decisions involve extensive discussions on the array of issues, including the program structure and its administration (who approves study designs, who pays for it, is the focus site-specific or regional, etc.), interpretation (what is an effect, what is unacceptable, will results be combined or interpreted individually, etc.), program objectives (what is the focus, the purpose, is research allowed, etc.), study design (is there a baseline, are there good reference sites, how much power and confidence do you want, how big a change do you want to be able to detect, how quickly do you need a decision, etc.), and data analysis (who analyzes the data, is the data public, how fast do analyses have to be completed, etc.).

Reaching consensus across stakeholders involves the development of trust, an atmosphere of transparency, and a commitment to working together. While a logical flow requires some questions to be answered before others, such as identifying goals and selecting indicators [7], addressing all the questions identified here (Table 1) seldom occurs in order. Regardless, all of the questions require resolution to achieve a design that is acceptable to all groups. Successful approaches used in other programs may also be transferred to the new program. In addition, the approaches may not be static and some approaches may not be feasible or achievable. The following overview describes many of the pitfalls, misalignments, and points of friction that are encountered in addressing these questions, and is meant to provide some guidance for scientists looking to enter into the development of a large regional aquatic monitoring and assessment program in a multi-stakeholder environment.

There is a lot of recent attention given to monitoring in terms of philosophy [8], in terms of enabling communities to participate and design programs [9], and integrating monitoring programs [10]. There is also differing attention to subtle differences between monitoring, modelling, measurement, and research, which will not be dealt with here. The focus of this paper is on the design of the monitoring program: the consequences of decisions around structure, objectives, design, reporting, and interpretation. The consequences of these decisions apply to all monitoring programs, regardless of whether they are designed by communities, by governments, by industry, or by environmental groups. It is the consequences of frictions in philosophies across groups that challenges multi-stakeholder groups to align in the development and design of the aquatic environmental monitoring programs.

2. Essential First Steps

Not surprisingly, as one enters the process, the first challenge is to align clarity around the objective of the program. Monitoring, while its objectives can be tightly defined and the questions narrowly constrained [7] can be widely interpreted. Friction is often generated by a lack of a shared vision around the objectives of the program and what data collection activities are or are not monitoring. A failure to reach agreement on what the monitoring is meant to achieve and how to achieve can become problematic as the program matures [7].

Multiple activities are considered for monitoring. A new program may be seen as an opportunity to get data related to the needs of the program, get data that there has been a desire to collect that may be related to the program, and an opportunity to collect other data related to the region that there has not previously been necessary funding or triggers to collect. While each may be necessary, each of these tasks can be distinct, but may be conflated under an overarching (and potentially inappropriate term) such as monitoring. Such conflation can be confusing and may lead to unnecessary friction.

An essential step to begin is to develop a specific, clear, and concise vision, objective, aspiration, purpose, and scope of the monitoring (see examples in Table 2). The discussions required to reach consensus on the purpose and objectives should not be underestimated. It is essential that there is broad agreement on the objective, purpose and scope, and that any "scope creep" is intentional, directional, and focused, and based on results coming from the program. Similar to the definition of a question (See Section 2.1 and Table 3), if balance between simple and a complex program are not attained, the program may drift toward a facile or grandiose program, respectively.

Table 2. Example of initial needs for program design. The example below is for a major event, and is meant as an example of the specificity required.

Vision	Any effects of an event are detected, tracked, and mitigated, and recovery tracked and documented.
Objective	To be able to detect change and predict effects, and adaptively manage for changing environmental conditions, with a science-based, integrated and transparent monitoring program.
Aspiration	Monitoring will identify the magnitude and extent of changes associated with an event or development (and how it is changing over time), and effects on receptors that can be associated with the event.
Purpose	To understand the impacts of an event and to develop a better understanding of the variability in responses in the system.
Scope	To detect site-specific, local, and regional change in areas of impacted by the event.

Table 3. Common problems with monitoring programs that represent the major failures in programs, in order of priority.

1 Ask too complicated or an unanswerable question
2 Not know what the answer is that you are looking for
3 Decide what endpoints are important after you start
4 Let non-scientific factors decide the approach and design
5 Figure out how big a change represents a concern after you start
6 Decide how many samples you need after you finish collecting
7 Assuming one study design can answer any possible question
8 Assume your data is good when you get it
9 Fail to adapt to your design or approach as you learn more
10 Assume that you understand what is going on

Once the scope has been developed and agreed upon, it is essential to make some decisions about the philosophical approach to the program. There are three independent philosophies to designing questions:

- Values-based questions, that deal with perception issues, such as "has the taste or risk from eating fish or shellfish changed?"

- Stressor-based questions that deal primarily with exposure issues and environmental impact assessment (EIA) predictions, such as "how far did the contaminants travel?"
- Effects-based questions that deal with the accumulated environmental state, such as "are there residual environmental concerns?" or "are there regional cumulative effects?"

Good monitoring programs need to combine the different philosophical approaches depending on the questions being addressed. Monitoring endpoints need to be a combination of measures that are ecologically relevant and measures that provide a signal before "damage" is irreversible (effects-based), as well as ones that enable you to determine cause, and measures that give you "early warning" (stressor-based), and also include measures that deal with public concerns (values-based) [11,12].

Each approach has embedded assumptions and common practices. Stressor-based questions tend to focus more on the earliest-warning indicators: measures that respond more rapidly, are more reversible, and are easier to relate to a cause. Stressor-based designs have other advantages; measurement techniques may be well-established and large datasets can be accumulated quickly. Stressor-based designs, however, also have disadvantages. In some cases no standard measurement techniques or reference materials may be available, such as 'naphthenic acids' in oil sands process-affected waters [13,14]. Safe exposure levels or guidelines may also not be available or may not be applicable. The accumulation of a chemical contaminant is not an impact but it is a stressor and presents a risk of an impact. Programs can become preoccupied with stressors [15] and with understanding mechanisms for stressors which don't appear to be affecting the biota.

Stressor-based designs have other attributes. In a site-specific industry monitoring program, questions originate primarily from controlling known stressors, and documenting expected or unexpected effects. Controlling known stressors is addressed with a stressor-based design intended to document change associated primarily with known stressors and exposure-response pathways. In general, the stressor-based approach requires data to address broad, basic, and initial questions about known drivers, stressors, and pathways. In addition to regional stressors such as climate change, fire, and invasive species, key drivers associated with activities need to be identified in a conceptual model to focus the description of key stressors. Stressor-based designs equate control of stressors with control of impact. While this approach has merit where and when stressors are clearly problematic, such as dioxin, pure stressor-based designs are sensitive to over-intervention and Type I errors [11].

Documenting unexpected effects or the relevance of changes in stressors is addressed with an effects-based design [12]. Effects-based questions tend to focus on biological endpoints. The measurements are ecologically relevant, but may have a long time lag between exposure and emergence of a detectable change, are harder to reverse, and more difficult to link to a cause.

Finally, values-based questions are the most difficult to handle, as indicators and tools are often more linked to the social science realm, or from a non-aboriginal perspective require translation to western science endpoints to effectively understand the causes and implications of changes. The translation of the questions involves significant challenges with both operationalizing them as well as their design and communication. This remains a substantial challenge for aquatic monitoring in Canada.

Once the philosophical basis for the program has been established, the next initial steps include defining the question, defining what an answer looks like, and deciding what the best indicators are for the desired answer. This description will deal with the important predesign questions at a philosophical level. There are other key, more detailed steps to adjust and adapt the program after the program has started, but they won't be dealt with in detail here. While trying to describe all of the steps to develop a program can be daunting, it is easier to conceptualize the characteristics of unsuccessful programs (Table 3).

2.1. Defining the Question

The main challenge affecting the design of many past programs is a failure to clearly define the question, to agree on the question, or the tendency to ask too complicated a question (Table 3).

Many monitoring programs suffer from trying to answer too many questions with a single approach, and it is important to understand that different components of the program will be required to answer different types of questions. While answers to each may be necessary and valuable, questions about the bioavailability of a contaminant, potential human health effects, or ecological consequences of development require different designs and approaches, needing multiple approaches should not be viewed as a deficiency. The key is to make the questions simple, agree on them, and design the monitoring to clearly answer the questions (e.g., are there regional cumulative effects; is stress increasing in the system; are unacceptable changes happening?). Data that is not contributing to key questions is not relevant to the program.

One of the points of friction in developing monitoring programs is the role of research-based questions, and the reluctance to adapt a program as more information develops (Table 1). Challenges arise when research is needed to further develop the monitoring tool, or the data collection is not targeted toward finding a specific answer. Curiosity-driven research should not be funded within a monitoring program. If the endpoints and targets are uncertain, research should be done to define the relevant questions and endpoints outside of the monitoring program per se, and there needs to be a research pot to provide funding for these questions.

Designing a successful program depends on what kinds of an answer is desired, and how tightly you can word the question to obtain the answer. Often the temptation is to insert a complex (and potentially unanswerable) question originating from policy or from social desires may be difficult to answer, such as "is mitigation necessary?", "has recovery occurred?", or "what are the cumulative effects of industry?". It is very difficult to get consensus between stakeholders on how to scientifically answer a question like "Is there a problem?" While defining areas of concern and focusing appropriate mitigation when necessary need to be part of the goals of a regional monitoring program, they are difficult to answer with a single monitoring effort and are more effectively answered with an iterative collection of data that builds the knowledge necessary to make management decisions. The focus of questions should be to define a situation where a monitoring response represents a signal that identifies a situation you want to know more about. In addition, while some questions may be salient and socially desirable, other are affected by technological capacity, safety, missing baseline data, and other issues (Table 1).

There are a number of principles used in the development of the Canadian Environmental Effects Monitoring (EEM) program [2,16] to address these often complex monitoring questions. EEM is a cyclical program with tiers of monitoring that include surveillance, confirmation, extent and magnitude, investigation of cause, and investigation of solutions. It is more effective in many cases to iteratively develop an understanding than to try and do it over a short time period, changing the timing and intensity of monitoring with increasing concern. As understanding increases, the design of specific components can and will improve.

The EEM program has been operating since the early 1990s [1], and long term elements of natural variability will create noise that can challenge the interpretation of ecological data (see [17]). Targets that satisfy the adequacy of data to answer the question need to be defined. The consequences of a measure failing to detect a response should not be that you need to look harder. It is as important to define what represents the absence of an effect than what is an effect.

The primary importance is to get the question scientifically and administratively right. It helps to subdivide complex questions, such as "is there a problem" into sub- (and serial-) questions that are easier to develop an answer to, including:

- Is there a difference (a statistical difference)?
- Is there a change (is the difference large enough that it surpasses a trigger that reflects natural variability)?
- Is it real (requires confirmation of change)?
- Was it expected (requires an understanding of risk and expected risks)?
- Is it stable or getting worse (requires temporal data)?

- How big an area is changing (requires understanding the extent and magnitude of change)?
- Is it meaningful (requires understanding how important a change is relative to other sites and other indicators)?
- Where might it be coming from (requires identifying the cause)?
- How serious is it and do I need to fix it or stop it (requires an understanding of ecological relevance)?

All of these need to be dealt with within a consistent design and operational framework with linked monitoring approaches and indicators interpreted within a broader framework. It bears repeating that the above questions cannot be answered within a single monitoring effort, and requires long(er)-term effort and program designs. For example, causes of observed changes may not be easily identified retrospectively and are limited to the data available. Results of these analyses are hypotheses, but they may nevertheless sometimes be treated as definitive conclusions. A potential response to these restrictions are to collect more data during the surveillance phase in hopes it may be useful in the future, but this strategy does not successfully address the "data first, questions later" criticism [18].

2.2. Defining Answers

In parallel to defining questions, early in the design phase, decisions have to be made to develop a clear understanding of what the target is and what a response will be. For water or air quality indicators, it is usually easier to develop targets, especially if there are existing environmental quality standards or guidelines. The common use of surrogates in monitoring mean defined targets reflecting when a change is unacceptable have not been developed. It is critical in the design of the program to consider the ability to detect changes (the power and confidence of indicators), and to collect the information required to inform a decision on the acceptability of a change. The relationship between the surrogate (measurement) and the assessment endpoint also requires consideration [19].

Making management decisions requires the development of a level of understanding that allows you to determine whether the change is expected and whether it is sustainable, and acceptable? Sustainability is, at its base, a question related to degradation in the environment, whereas acceptability has to do with decisions about restoration to a past or preferred state. It is important to separate the questions of sustainability and acceptability, and the priority has to be to be able to detect degradation, since mitigation of impacts while the situation is still degrading can be challenging. Having lost 25% of the species can be sustainable, but the question of acceptability means that it may not be tolerable. Separating the questions allows progress without fully understanding the historical state—understanding whether degradation that continues to occur can focus assessments. Relationships between the sustainability and acceptability of changes are not commonly explicitly considered or defined in aquatic monitoring programs.

In order to effectively manage the environment, you need to be able to measure responses to the stressors or you are not effectively tracking development. One of the stumbling blocks to move forward in environmental management has been the hesitancy to measure a change for fear that someone would be forced to fix it. At some level, with sensitive enough tools, you will be able to measure environmental change associated with development. This challenge becomes more acute as technological sophistication advances and smaller and smaller quantities of chemicals can be measured [20]. Understanding the ecological relevance of a change, in other words the sustainability and the acceptability of that change, that is relevant for decision-making and counterbalances the potential threat of over-intervening.

The focus of monitoring should be on detecting changes, especially changes that are getting worse, before they become ecologically relevant, and while there is time to limit any potential damage. It is easier to define what an answer looks like if you look at the sub-questions, than if you try to define the answer based on whether or not there *is* a problem, and it helps to understand that different sets of data will be required to answer different questions, and the study design needs to incorporate the ability to tier the analyses. For example:

- Is there a difference: involves comparing the data from a site of interest to relevant local reference site(s);
- Is there a change: is a question asked over time at a single site, in comparison with available historical or reference data;
- How big an area: involves a spatial data set;
- Is it getting better or worse: involves a spatial data set over time;
- How serious is it: requires ecologically relevant endpoints across a range of reference sites.

The key is to set a target that reflects the level of concern, and it is critical to set the target at a level that demonstrates acceptable environmental quality to reduce the tendency to look harder when effects are not present.

3. Designing an Adaptive System

An important indicator of a "world-class" monitoring program is that it be "adaptive and robust: an approach that can be evaluated and revised as new knowledge, needs, and circumstances change and that ensures stable and sufficient funding" [6]. It is important to develop a framework for an adaptive monitoring strategy, and to develop triggers to improve your sensitivity and monitoring program management responses [10,21,22] (Table 1). Many of the traditional monitoring endpoints do not have the sensitivity and response times necessary to allow the program to respond in a proactive, protective fashion. Early decisions are needed for determining the key parameters that would signify a change in monitoring is needed, and the level of change that would be required to change the management strategy. Power analysis can then be conducted to determine whether the sampling intensity is sufficient to allow protection.

In an adaptive monitoring cycle, monitoring design, emphasis, and focus evolves through feedback loops that accelerate or decelerate activity as necessary, and includes other components to improve the program over time [10,11,22]. Adaptive monitoring is composed of hypothesis-driven questions linked to conceptual effect pathways, via tiers of monitoring. Movement between tiers is governed by triggers and reflects adjustment of intensity, frequency, and focus of monitoring to channel effort where there is the most concern.

The Canadian EEM program provides one model of an adaptive monitoring program [1,2]; a similar model has been developed in Brazil [5]. Monitoring questions can be divided into long-term, focused, and baseline studies. Long-term (surveillance) studies should form the core of the monitoring program. They integrate adaptive, but bounded and prescriptive steps (tiers) guided by trigger exceedances. They provide consistent monitoring of sites and indicators over a time period typically >5 years to evaluate the state of the environment. Consistency of measurements and sites is intended to address multiple interpretative objectives, including answering site-specific, local, and regional questions. Periodic adaptation of the long-term monitoring network is required to ensure it achieves the stated monitoring objectives, but also as a mechanism to adopt new information as more is learned. In long term programs, regular sampling of sites would typically occur every 3 or 4 years for endpoints other than routine chemistry.

Focused studies (extent and magnitude) are typically a 2–3 year monitoring/research spatially limited activity usually triggered by signals generated by the long-term monitoring network and designed to answer a specific monitoring question related to evaluating the extent and magnitude of a change of concern, or to address specific knowledge gaps about system processes or function. Focused studies are also used for special (and possibly stand-alone) topics, including feasibility of candidate monitoring tools or to resolve an unanticipated issue.

Baseline studies primarily deal with meeting minimum baseline data requirements (≥3 years) to more effectively evaluate the occurrence of change during the long-term studies and in areas that will be developed further in the future or have been exposed to an event. It is important to note that "baseline" does not need to refer to a historical condition. Since sustainability is an issue of degradation, the current state can be used as an indicator to track progress or change. Historical state does need to

be estimated to help assess acceptability and recovery, but the absence of historical data should not be used as an excuse to limit monitoring.

There are additional types of studies that may be triggered via long-term, focused, or baseline work:

- Confirmation studies: a repeat of surveillance monitoring, on an accelerated pace, to evaluate the replicability of a change.
- Investigation of cause (IOC): are generally hypothesis-driven studies designed to characterize the potential cause or source of an issue after evidence that the change is real and importance is obtained. IOC is a type of focused study.

All of the components and approaches need to be integrated into an overall approach (Table 4).

Table 4. A breakdown of components of an idealized regional monitoring program.

1. Monitoring exposures (stressor-based)
   a. Source (routine monitoring by industry)
   b. Ambient (baseline and enhanced stations; water includes quality and quantity stations, tributary, mainstem and lake/ocean)
   c. Deposition
   d. Receptor exposures
2. Monitoring ecosystem effects (effects-based)
   a. Component (community and/or sentinel species components; for water need mainstem, tributary and lakes/ocean)
3. Stakeholder-driven monitoring issues (values-based; triggered by regional stakeholder concern)
4. Cross-components
   a. Background/baseline
   b. Baseline data for new sites
   c. Supporting data (meteorological, hydrological)
5. Focused monitoring (triggered in by a specific concern)
   a. Extent and magnitude of effects or responses
      i. Spatial distribution
      ii. Temporal distribution
      iii. Change from historical
   b. Examination in alternate species, approaches or levels of organization
6. Research (prioritized data gaps)
   a. Investigation of cause (when effects are seen)
      i. Characterizing exposures/ sources
      ii. Ecological consequences of impacts
      iii. Management implications
   b. Modelling
      i. Development—improving data, equations or testing assumptions
      ii. Validation—data integration, confirmation or calibration
      iii. Estimating predicted or historical trends
7. Methods development fingerprinting, new measurement validation or remote sensing

3.1. Development of Triggers

Identifying unexpected change and confirming expected changes are occurring is a purpose of monitoring. Monitoring triggers can be designed to initiate more detailed studies (tiers) at specific locations, more locations, or using more detailed questions or information from additional indicators, etc. Observations that are "different than expected" will be flagged by the definition of expected normal ranges, or triggers [21,22]. Triggers can be developed from existing data where adequate, or from the published literature as an interim trigger if local data are not available. They can also be based on time-since-last sampling (time triggers aka regular site rotation), applied randomly where neither data nor time triggers have been exceeded (in a surveillance/confirmation mode), or triggered in by findings in another component (i.e., changes in water quality). To develop a baseline, a minimum of three years of data are needed; during these baseline years, spatial comparisons predominate. Following the baseline collection period, test years will work in a surveillance mode and focus on testing for change within a location, locally and regionally.

When change exists, it is a priority to address whether the change was expected or unexpected, and whether it is stable or getting worse. When changes are higher than expected (exceed a trigger), confirmation monitoring is triggered, and then focused monitoring to examine the extent and magnitude

are conducted. If change is of sufficient concern, studies can proceed directly to the investigation of the cause.

3.2. Defining How Big a Difference Is Going to Be Interpreted as a Response or Signal

The ability to make predictions and an awareness of the magnitude of difference that should be a warning signal are important tools for risk assessment and management. These warning signals should be used to adapt the monitoring program, including turning off and on monitoring stations when the signals warrant a change in strategy, and prompting site-specific evaluations when ecologically significant changes are occurring. At new developments, an increase in seriousness of lower level changes can be predictive of future potential higher level concerns.

A critical component of a monitoring program must be to develop the baseline understanding of how much variability exists in measurements, and how big a difference will signify a *meaningful* change. Natural variability is one of the major concerns. Noise in field measurements arises because of several factors, including natural variability in environmental factors, the adequacy of reference data, the linkage of the indicator to meaningful performance, the presence of confounding factors other than the stressors you are interested in, and measurement and sampling error. There is a lot of confusion between conflicting and overlapping issues that affect interpretation such as whether the reference site you are using is valid, how much of the noise in the data is natural variability, and whether changes are ecologically relevant. The key things to move toward an answer is deciding what is a reference to compare to, how much natural variability is there (and how do I reduce the noise), and how big a difference is ecologically relevant.

Any monitoring program needs to be iterative, using data to adapt the program and improve the interpretation of the outputs. The most effective manner is to develop triggers that allow the program to adapt, including for turning on and off stations or measurements, for establishing new stations, for evaluating the adequacy of measurements and for changing the frequency of monitoring. The program should be designed to adapt and evolve while maintaining the integrity of the monitoring system, stations, and measurements. Program tiers can include surveillance, confirmation, focused monitoring, and investigation of cause, and each level should have triggers (Table 5).

Table 5. Examples of tiers (or phases) of monitoring, similar to those used in environmental effects monitoring.

Tier	Example Trigger	Question	Frequency
Basic		Are there changes?	Regular
Confirmation	Difference beyond a critical effect size threshold (natural variability)	Can we confirm them?	More often
Extent	Confirmation of changes (reference site adequacy)	What is the extent and magnitude of the change?	More stations and indicators
Cause	Change across a sufficient area or of a sufficient magnitude, or is getting worse (temporal consistency)	What is the cause?	Research-oriented
Concern	Change exceeds "ecological relevance"	What is the solution and do I have to mitigate or compensate?	Hopefully never

Responses that surpass a pre-designed target response level act as a warning signal that monitoring needs to increase to develop an assessment process. Trigger development needs to focus on defining the size of a change that will mean that I am concerned enough to increase monitoring or enter assessment or investigation of cause. Defining these triggers is critical to understanding what an answer to the questions will look like. The focus of the trigger needs to be more sensitive than a trigger that warrants a management response in terms of mitigation or correction.

During the early stages of sampling, an enhanced monitoring program is needed to develop the triggers and the background information to interpret potential future changes, should enhanced monitoring be required. As the triggers are developed, sufficient information will become available

that sites or sampling times become redundant, and sites or times can be triggered off to reduce monitoring costs, as long as parameters at key sites stay within defined levels (below triggers). Values that exceed trigger levels turn back on some of these sites or sampling frequencies to understand the extent and magnitude of the deviation, and to determine the significance of deviations.

4. How Do I Pick the Right Indicators?

There are opposing camps in the philosophy of picking indicators from those that want to "capture the variability" and those that want to "focus the assessment." All decisions about indicators (endpoints) are a forced compromise between decisions about the relative importance of the ecological relevance, the tolerance for a time lag, the desire to want to know cause, and the reversibility of a change. Changes at the community level (biodiversity) are hard to reverse, take a long time to develop, and are difficult to link to a cause, but they are highly relevant and the ones that best resonate with decision-making. At the other extreme, changes in molecular or physiological endpoints occur quickly, are easy to reverse and easier to link to a cause, but are not very relevant ecologically. Monitoring at the community level accepts that there are many easier to reverse changes that would have been detectable earlier than you would detect a community change. Monitoring at the physiological level means that there are many changes you can detect that may not translate into effects at higher levels of organization.

Traditional monitoring approaches have been very bad at anticipating potential issues, have lacked the sensitivity to be predictive, and have often focused either on upper level community and biodiversity issues that do not allow sufficient time lag to correct changes before they may become irreversible, or totally on chemistry and biomarkers that lack the information on ecological relevance to allow decisions about acceptability. Biomarkers are a good tools to confirm exposure of fish, and residency, but if they don't impact higher functions (growth, survival, reproduction, energy storage) then they aren't meaningful other than as exposed. If they do impact higher function it is easier, cheaper and faster to measure the whole organism endpoints.

Monitoring programs need to have overlap in indicators and need to include three types of indicators:

- Early warning indicators that tell you whether predicted or anticipated changes are happening; they are usually at lower levels of organization, or direct measures of the stressors of interest;
- Performance (effects) indicators that are integrators that tell you whether the accumulation of stress is affecting indicators that threaten sustainability;
- Biodiversity type indicators that tell you whether changes at lower level are important enough that damage has been done.

The relevant balance of indicators will vary with how well you understand your situation, how much warning is needed, and how much tolerance there is for time lags, and in general how much protection is warranted. The key concept is to define the target that will mean the environment is protected. There are two types of protection you need to worry about—protection from finding a response that is not real (type I statistical error), and protection from not finding a real response when it exists (type II statistical error). In the first case, an adaptive monitoring program responds to finding an "effect" (a change greater than a critical effect size) by triggering confirmation steps, which include changing the frequency of monitoring, and potentially the extent and magnitude of sampling. If the effect is not real, the program can revert to its previous monitoring state and intensity, and there is enhanced protection from false positives. So type I errors (finding a response that was not real) should not be a major concern for monitoring, as long as there are conservative requirements for confirmation in case an effect is detected.

In the second case, the power of any design is determined by the probability of making a false declaration of an effect. Power is one of the factors involved in determining the number of samples (or years) required for developing the information required to make decisions, but the driving force is

the variability of the measurement endpoint. Other factors affecting power (sample size, critical effect size, and statistical criteria) can be easily adjusted but it takes more effort to reduce the variability and retain sensitivity. It is also the variability in the endpoint between rounds of cycling at a site which is not changing (the natural variability), that drives both your ability to detect an effect and the size of a critical effect size that can cost-effectively trigger a change in monitoring strategy. The time lag before a detectable response occurs also has a major impact on the sensitivity and responsiveness of a program.

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. False negatives are the real thing to worry about because over time changes which were undetected have the potential to translate into broader and more significant changes if they persist. A type II error (not finding a real effect) is the main issue that stakeholders worry about—there is concern if you didn't find something that you missed it because you weren't looking hard enough. So the only solution to avoiding type II errors is to increase your power (by increasing your sample size and reducing your variability) so much that you increase your chance of type I errors. Simply, the only solution to missing an effect is to find effects that are not real using a very sensitive monitoring program. Therefore, the response to any determination of effect has to be tempered, and include steps for confirmation, and extent and magnitude before evaluating the significance of a change. As mentioned earlier, with sensitive enough tools it is possible to measure change which is both sustainable and acceptable.

It is important for the adaptive monitoring strategy to focus on changes that mean protection, and that achieving good performance in those levels gives comfort that management changes are not warranted. The critical factors to consider in choosing the types of questions that are needed are:

- How confident am I if I don't see a change in my measurement endpoints that nothing important is happening (am I monitoring at the right level?);
- If I do detect an effect, where is the next obvious place I would look to evaluate how important it is (where else do I need baseline data?);
- What is the variability in the measurement endpoints (within and between cycles) and how much power does that give my study, given the sample sizes and frequency of monitoring that I can afford (how much statistical power do I have?);
- How big a change in these endpoints should create a situation where I want to know more information (what is the critical effect size?); and
- How confident am I that a change in these endpoints gives me concern that specific operation is having a potential impact (what is the potential that I can link changes in these endpoints back to specific operations?).

The indicators have to address a clear question that is relevant to the objectives of the program, and have to have an answer that is definable, and that is relevant to the program. Asking scientists what they want to measure is not as important as asking the government and industry, under which conditions (what results) would make them change how they manage the system or their operations. One of the common problems in developing a monitoring program is the tendency for everybody to want to include their measure, or redundant measures, or pet measures, because everybody feels that their measurement is the "best." In general, all scientists want to measure what they are good at and what they have experience with, so there needs to be some real input into indicator selection from decision-makers. The important question is what level of change will be needed to make decisions. The Canadian EEM program focused on benthic invertebrates and fish, with supporting information from chemistry and toxicity tests, because those are the levels at which decisions are usually made under Canadian legislation. The challenge with zooplankton, phytoplankton, and biomarkers is that if you see effects, you (at least in Canada) would have to trigger in further study to determine the ecological relevance of the changes (unless you are going to regulate based on zooplankton changes if they don't result in changes to fish or benthos).

When there are conflicts or disagreement about what measures should be used, a series of tiered criteria can be developed to screen endpoints for their relevance to the program. Once the key questions have been developed, and decisions made as to what the answers need to look like to make decisions, the screening criteria can be applied to help make decisions about the suitability, applicability, and acceptability of proposed indicators. While there are an abundance of criteria for selecting indicators, most of the long lists of criteria are biased and self-serving. The Canadian EEM program settled on a relatively simple list (sentinels needed to be abundant, exposed and the indicators needed to be measurable).

Other important criteria are balanced, relatively easy to measure (cost-effective), not redundant, and will trigger a concern when relevant change is present. During the development of the federal EEM program, the kinds of criteria used to evaluate measures included evaluations of whether the measures:

- Answer a clear relevant question;
- Have a published, peer-reviewed analytical protocol;
- Are commercially available;
- Have an adequate baseline or reference for comparison to (what would a change be?);
- Have defined targets and interpretation process (what would happen if threshold exceeded; i.e., when does a change occur?).

For measures such as chemical analytical methods, there were additional criteria, including issues such as whether there has been a valid interlaboratory comparison for analytical endpoints, and whether the detection limits known, specified, adequate, and achievable and whether adequate analytical standards are available. There are other additional criteria that can be used, but would need to be developed for any specific program.

Once the indicators are developed, there are additional steps including deciding where and when to sample to get the most sensitive indicators [23], determining how big a difference is going to be interpreted as a response [22,24], and determining how many samples you will need to detect the effect.

5. Final Considerations

Regardless of the effort put into designing a program, there are some common weaknesses that are difficult to design around, including concerns about the adequacy of reference sites, the presence of confounding factors, and the influence of natural variability [25]. There is an under-appreciation for the role of scientific consensus, replicating results, and other brakes in the process between research findings to management intervention (to stop us from making decisions based on flawed information). There are some key final considerations that need to be considered during the development of a monitoring program, including that most programs try to do too many different things at once. It is necessary to make the questions simple, and to have clear agreement on how data will be interpreted. There are trade-offs (strengths and weaknesses) in any monitoring program, and there is a lot of noise in data and you need to work to understand what drives natural variability to get to more sensitive monitoring programs. Finally, it will take considerable amounts of time to understand the influences of natural variability—answers do not come quickly.

Author Contributions: Conceptualization, K.R.M.; Methodology, K.R.M., T.J.A. and M.A.G.; Writing—original draft, K.R.M.; Writing—review & editing, T.J.A. and M.A.G.

Funding: This review received no external funding.

Acknowledgments: The contents of the review reflect more than 30 years of discussions among a wide variety of stakeholders, and we appreciate the many philosophical discussions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lowell, R.B.; Ribey, S.C.; Ellis, I.K.; Porter, E.; Culp, J.M.; Grapentine, L.C.; McMaster, M.E.; Munkittrick, K.R.; Scroggins, R.P. *National Assessment of the Pulp and Paper Environmental Effects Monitoring Data*; NWRI Report 03-521; National Water Research Institute: Burlington, ON, Canada, 2003.
2. Walker, S.L.; Hedley, K.; Porter, E. Pulp and paper environmental effects monitoring in Canada: An overview. *Water Qual. Res. J. Can.* **2002**, *37*, 7–19. [CrossRef]
3. Ribey, S.C.; Munkittrick, K.R.; McMaster, M.E.; Courtenay, S.; Langlois, C.; Munger, S.; Rosaasen, A.; Whitley, G. Development of a monitoring design for examining effects in wild fish associated with discharges from metal mines. *Water Qual. Res. J. Can.* **2002**, *37*, 229–249. [CrossRef]
4. Chiang, G.; Munkittrick, K.R.; Orrego, R.; Barra, R. Monitoring of the environmental effects of pulp mill discharges in Chilean rivers: Lessons learned and challenges. *Water Qual. Res. J. Can.* **2010**, *45*, 111–122. [CrossRef]
5. Furley, T.H.; Perônico, C. (Eds.) *Guia Técnico de Monitoramento dos Efeitos Ambientais em Corpos Hídricos*; Instituto Aplysia: Vitória, Brazil, 2015; 264p, ISBN 978-85-5642-000-8.
6. Environment Canada. *An Integrated Oil Sands Environment Monitoring Plan. Cat*; No. En14-49/2011E-PDF; 2011; ISBN 978-1-100-18939-0. Available online: http://publications.gc.ca/collections/collection_2011/ec/En14-47-2011-eng.pdf (accessed on 9 February 2019).
7. Downes, B.J.; Barmuta, L.A.; Fairweather, P.G.; Faith, D.P.; Keough, M.J.; Lake, P.S.; Mapstone, B.D.; Quinn, G.P. *Monitoring Ecological Impacts: Concepts and Practice in Flowing Waters*; Cambridge University Press: Cambridge, UK, 2002; ISBN 9780521065290.
8. Lindenmayer, D.B.; Likens, G.E. *Effective Ecological Monitoring*, 2nd ed.; CSIRO Publishing: Clayton, Australia, 2018; ISBN 9781486308927.
9. Conrad, C.C.; Hilchey, K.G. A review of citizen science and community-based environmental monitoring: Issues and opportunities. *Environ. Monitor. Assess.* **2011**, *176*, 273–291. [CrossRef] [PubMed]
10. Somers, K.M.; Kilgour, B.W.; Munkittrick, K.R.; Arciszewski, T.J. An adaptive environmental effects monitoring framework for assessing the influences of liquid effluents on benthos, water, and sediments in aquatic receiving environments. *Integr. Environ. Assess. Manag.* **2018**, *14*, 552–566. [CrossRef] [PubMed]
11. Arciszewski, T.J.; Munkittrick, K.R.; Scrimgeour, G.J.; Dubé, M.G.; Wrona, F.J.; Hazewinkel, R.R. Using adaptive processes and adverse outcome pathways to develop a meaningful, robust, and actionable environmental monitoring programs. *Integr. Environ. Assess. Manag.* **2017**, *13*, 877–891. [CrossRef] [PubMed]
12. Munkittrick, K.R.; McMaster, M.; Van Der Kraak, G.; Portt, C.; Gibbons, W.; Farwell, A.; Gray, M. *Development of Methods for Effects-Based Cumulative Effects Assessment Using Fish Populations: Moose River Project*; SETAC Press: Pensacola, FL, USA, 2000; pp. 236 + 18.
13. Tanna, R.N.; Redman, A.D.; Frank, R.A.; Arciszewski, T.J.; Zubot, W.A.; Wrona, F.J.; Brogly, J.A.; Munkittrick, K.R. Overview of existing science to inform oil sands process water release—A technical workshop summary. *Integr. Environ. Assess Manag.* **2019**, *15*, 519–527. [CrossRef] [PubMed]
14. Fennell, J.; Arciszewski, T.J. Current knowledge of seepage from oil sands tailings ponds and its environmental influence in northeastern Alberta. *Sci. Total Environ.* **2019**, *686*, 968–985. [CrossRef] [PubMed]
15. Lima, A.C.; Wrona, F.J. Multiple threats and stressors to the Athabasca River Basin: What do we know so far? *Sci. Total Environ.* **2019**, *649*, 640–651. [CrossRef] [PubMed]
16. Environment Canada. *Environmental Effects Monitoring Technical Guidance*; Environment Canada: Gatineau, QC, Canada, 2012. Available online: https://www.ec.gc.ca/esee-eem/default.asp?lang=En&n=AEC7C481-1 (accessed on 9 February 2019).
17. Bowron, L.K.; Munkittrick, K.R.; McMaster, M.E.; Tetreault, G.; Hewitt, L.M. Responses of white sucker (*Catostomus commersoni*) to 20 years of process and waste treatment changes at a bleached kraft pulp mill, and to mill shutdown. *Aquat. Toxicol.* **2009**, *95*, 117–132. [CrossRef] [PubMed]
18. Lindenmayer, D.B.; Likens, G.E. Adaptive monitoring: A new paradigm for long-term research and monitoring. *Trends Ecol. Evol.* **2009**, *24*, 482–486. [CrossRef] [PubMed]
19. Suter, G.W.; Vermeire, T.; Munns, W.R.; Sekizawa, J. Framework for the integration of health and ecological risk assessment. *Hum. Ecol. Risk Assess.* **2003**, *9*, 281–301. [CrossRef]

20. Munkittrick, K.R.; Arciszewski, T.J. Using normal ranges for interpreting results of monitoring and tiering to guide future work: A case study of increasing polycyclic aromatic compounds in lake sediments from the Cold Lake oil sands (Alberta, Canada) described in Korosi et al. (2016). *Environ. Pollut.* **2017**, *231*, 1215–1222. [CrossRef]
21. Arciszewski, T.J.; Munkittrick, K.R. Development of an adaptive monitoring framework for long-term programs: An example using indicators of fish health. *Integr. Environ. Assess. Manag.* **2015**, *11*, 701–718. [CrossRef]
22. Kilgour, B.W.; Somers, K.M.; Barrett, T.J.; Munkittrick, K.R.; Francis, A. Testing against "Normal" with environmental data. *Integr. Environ. Assess. Manag.* **2017**, *13*, 188–197. [CrossRef] [PubMed]
23. Barrett, T.J.; Munkittrick, K.R. Seasonal reproductive patterns and recommended sampling times for sentinel fish species used in environmental effects monitoring programs in Canada. *Environ. Rev.* **2010**, *18*, 115–135. [CrossRef]
24. Munkittrick, K.R.; Arens, C.J.; Lowell, R.B.; Kaminski, G.P. A review of potential methods for determining critical effect size for designing environmental monitoring programs. *Environ. Toxicol. Chem.* **2009**, *28*, 1361–1371. [CrossRef] [PubMed]
25. Munkittrick, K.R. Ubiquitous criticisms of ecological field studies. *Hum. Ecol. Risk. Assess.* **2009**, *15*, 1–4. [CrossRef]