REPORT ON:

Guidance for a Weight of Evidence Approach in Conducting Detailed Ecological Risk Assessments (DERA) in British Columbia

Submitted to:
The Ministry of Environment
October 2010

Submitted by:
Science Advisory Board
For Contaminated Sites in British Columbia

Prepared under contract by
Exponent Inc
Bellevue, Washington USA
June 2010

Acknowledgements

This report, *Guidance for a Weight of Evidence Approach in Conducting Detailed Ecological Risk Assessments (DERA) in British Columbia* including annotated literature review has been undertaken as a supplement to the Detailed Ecological Risk Assessment guidance submitted to the BC Ministry of Environment in September 2008. It is presented for the information and benefit of the Contaminated Sites community in British Columbia. It is hoped that it will be of interest to practitioners in other jurisdictions as well. Earlier DERA Report is also posted on the SABCS website.

The study was undertaken by Exponent Inc, Bellevue Washington. The capable and knowledgeable Exponent team was lead by Anne Fairbrother. The work was supervised by SABCS through a Task Force chaired by Gary Lawrence of Golder Associates Ltd. The personal involvement of Gary in liaison with Exponent and his interest in the field of study was appreciated. The Task Force included a number of recognized risk assessors in British Columbia who had previously served on the Risk Forum Weight of Evidence White Paper group:

- Beth Power, Azimuth Consulting
- Ryan Hill, Azimuth Consulting
- Blair Macdonald, Golder Associates
- Joline Widmeyer, EBA Engineering Consultants
- Doug Bright, AECOM
- Will Gaherty, Pottinger Gaherty Environmental Consultants Ltd. (SABCS Director/Board Liaison)

Additionally Remi Odense was the Ministry of Environment representative on the committee and Ute Pott of Environment Canada also served on the committee. Helpful input from Trish Miller, Senior Risk Assessment specialist at Golder Associates Ltd. in final review of the Report is also acknowledged. The SABCS much appreciates their insight, dedication. and patience in the development of weight of evidence guidance in British Columbia.

The SABCS acknowledges with appreciation grant funding from the government of British Columbia through the Ministry of Environment throughout the course of the project and the preceding DERA development. That funding made this work possible.

--

Science Advisory Board for Contaminated Sites in British Columbia
October 2010

Disclaimer

Practitioners and others with interests in contaminated sites should be aware that this report, including the appendix and other sections, intended as a supplement to DERA has not been adopted in whole or in part by the British Columbia Ministry of Environment at this time. This disclaimer also applies to the comments in the memorandum submitted to the SABCS by the Task Force which follows these acknowledgements. While every effort has been made to incorporate the best available science, these materials should be used solely as scientific review and commentary by the reader and applied in practice solely at the readers discretion and responsibility. This disclaimer is consistent with SABCS Policy.

Use of this Material

Readers are reminded that they are welcome to download a complete copy of the report and appendices for their personal technical and scientific use but that the reproduction of the work in whole or in part for commercial purposes or presentation can only be done with the express written permission of the Science Advisory Board for Contaminated Sites in British Columbia.

Request for Comment

The Science Advisory Board for Contaminated Sites in British Columbia is soliciting comment on the documents which together constitute a report to the BC Ministry of Environment on recommendations for the development of guidance on Weight of Evidence approaches in Detailed Ecological Risk assessment for practitioners in British Columbia. Comments will be reviewed and compiled by the SABCS, and will be much appreciated.

Please send your comments to the Science Advisory Board for contaminated Sites by email or email attachment to pwest@uvic.ca. Comments received by January 15,2011 will be most useful in further refinement of this work. However comments at any time on SABCS work are always appreciated.

--

Paul West, President Science Advisory Board for Contaminated Sites in British Columbia

MEMORANDUM

September 7, 2010

To: Paul West
Science Advisory Board for Contaminated Sites in British Columbia

From: Gary Lawrence, M.R.M., R.P. Bio.
Chair, WOE Task Force

Review by: Trish Miller,MSc. R.P.Bio.

RE: WEIGHT OF EVIDENCE (WOE) TASK FORCE COMMENT – FINAL WOE GUIDANCE

On behalf of the Weight-of-Evidence Task Force, I am pleased to provide final comment and advice concerning the final version of the deliverables from Exponent (including the Appendix IV titled Weight of Evidence Approach) and the related inserts for the Detailed Ecological Risk Assessment technical guidance.

INTRODUCTION

Because the Task Force assisted with framing some of the issues explored in the guidance (by way of Terms of Reference, multi-stakeholder meetings/conferences, and review of draft materials) we do not find it necessary to provide detailed comment on the content of the deliverables. Instead, in this memorandum we have emphasized the provision of advice to the Science Advisory Board for Contaminated Sites in British Columbia (SABCS) on the document, particularly in terms of how we believe that the guidance should be implemented. This advice is intended to supplement the Exponent submission, with emphasis on the following objectives:

- Provide general advice on interpretation/implementation;
- Identify aspects of the guidance that are most likely to change over time;
- Identify policy issues that BC Ministry of Environment (MOE) may consider; and,
- Identify components that would benefit from enhancement should resources come available or where the science is most active.

Our understanding is that both the SABCS and the Ministry of Environment appreciate this approach, and that this memo will be circulated among the SABCS Board members for comment prior to submitting the WOE guidance material to the Ministry.

INTERPRETATION

Overall, the Task Force was pleased with the final deliverables and believe that the guidance provides an important and worthwhile enhancement to the detailed ecological risk assessment (DERA) guidance. Provided that the guidance is

appropriately implemented, it will provide risk assessment practitioners with a defensible framework for conducting WOE assessments in BC. Importantly, the framework presented is suitable for application to a range of site conditions (from relatively simple to complex), a range of ecosystem types, and differing experience levels of risk assessors.

Early in the development of the Terms of Reference, the Task Force discussed the desired level of complexity, prescription, and quantitative detail that would be appropriate for WOE guidance. In the course of their literature review, Exponent identified a range of approaches that have been applied by others, and have recommended a procedure that is intermediate along the continuum of potential approaches. The Task Force believes that such an approach provides a good starting point for many sites. Although alternative methodologies are acceptable (and remain applicable within the broad WOE framework), the examples provided are a useful baseline for the application of WOE.

Two distinct advantages of Exponent's example approach are that it: (a) provides a clear template to non-specialists; and (b) offers a generally standard approach that should be applicable to most BC risk assessments, potentially simplifying the review process. Although several members of the Task Force have emphasized the need for flexibility in implementation, the guidance provides language that clearly states that the provided document describes a default weighting approach (and that alternative approaches may be used). Accordingly, we endorse the guidance, subject to the following directions for interpreting and appropriately applying the WOE guidance:

- Users should read the whole document, and consider the statements regarding flexibility to use alternative approaches. They should also consider the advantages of the "example" approach in terms of streamlining the review process. Here, we use the terms "example" and "default" interchangeably, meaning that the Exponent procedure will satisfy the technical and reporting needs for most sites, without necessarily being preferable to alternative approaches. Use of the term "default", therefore, does not equate with an expectation that risk assessments must follow that approach.

- The reader should understand that some details of the Exponent methodology or specific examples may not be appropriate for all risk assessments (whereas the broader concepts will remain applicable).

- Alternative WOE approaches may be used provided that the details of the proposed procedures are clearly explained (e.g., in terms of the specific attributes used to evaluate relative merits of lines of evidence (LOE), degree of quantification of these attributes, formality of decision rules for ranking or scoring of attributes). In general, the greater the degree of

departure from the default procedure, the greater the onus on the practitioner to explain the logic behind the alternative approach. The Task Force felt strongly that transparency was an important element of any WOE approach; as such, clarity in the WOE procedure is mandatory regardless of the methodological details. Because the default weighting approach provides a built-in logic system for aggregating multiple lines of evidence, alternative approaches will require clear explanation.

- Alternative approaches, where applied, should address the major conceptual issues raised in the default WOE procedure, as these high-level issues discussed in the guidance are considered to be universally applicable. For example, with respect to the weighting of attributes during Problem Formulation (PF) stage, it may not be strictly necessary to apply an a priori numerically-weighted approach to assigning values. However, should an investigator propose an alternative approach, it would not be acceptable to dismiss (skip over) the general requirement to give advance consideration of the merits and uncertainties of candidate lines of evidence (particularly during PF stage).

- The WOE framework links strongly to uncertainty assessment, which is an important aspect of all risk assessments, including those that rely on a single line of evidence. For example, a screening-level wildlife risk assessment frequently involves a single LOE (e.g., a food chain model relative to a literature-based benchmark dose). Although the uncertainty associated with the food chain model itself is often well-described in a risk assessment, it may be incomplete if it does not contrast that level of certainty to other risk assessments that benefit from a diversity of LOE (particularly those that include site-specific toxicity, bioavailability, or resident community measures). A common currency for describing uncertainty enhances transparency.

FUTURE CHANGES

The development of the Exponent WOE guidance recognized the range of approaches currently available for WOE, and several alternative systems were considered based on their literature review findings. The state of practice in ecological risk assessment (ERA) is constantly evolving and we believe that this warrants a combination of good scientific principles (i.e., as per the Exponent example) with ongoing flexibility in implementation. Both Task Force members and the Exponent team have attended recent technical workshops and conferences for which the topic of WOE has been emphasized. Although the details of the emerging methods and approaches change over time, a consistent refrain has been the promotion of a procedure that is rigorous, transparent and technically sound. The high-level concepts embodied in the Exponent guidance are therefore unlikely to become stale in the near future.
Some of the procedural details of the Exponent examples may change over time

as the field of risk assessment evolves in BC. Some of the specific areas for which change may occur most quickly are:

- Decision rules for weighting attributes – The scoring system for weighting of attributes (Table 1 Appendix IV) requires clear boundaries between the five attribute scores. The thresholds between categories shown in the example may not be applicable and/or may change over time. For example, the effect sizes articulated under "sensitivity and specificity" could change depending on the context of the ERA and/or on regulatory considerations (such as whether the receptor of interest is an endangered species);

- Magnitude definition – As discussed in our Workshop and as articulated in the WOE Guidance, the "magnitude of response" can be interpreted in many ways, including consideration of numerical magnitude, spatial scale, level of ecological organization, permanence, probability or frequency of occurrence, etc. Whereas the guidance provides some suggestions for characterizing the magnitude of response, the interpretation of risk magnitude is an area of active development in the risk assessment field.

The above concepts are closely linked to policy development (discussed further below). Both provincial and federal jurisdictions have been exploring opportunities to harmonize decision rules for concepts such as "acceptable effect sizes." Again, we emphasize that the broad concepts in the procedure (consideration of attribute values using a systematic process) are more important than the methodological details.

POLICY AND REGULATORY ISSUES

The WOE guidance has not been developed in a vacuum, but rather links to the existing and emerging regulatory regime. The most obvious linkage is to the DERA technical guidance, to which the WOE guidance is to be appended. It is our understanding that the Ministry intends to adopt DERA (including WOE) by developing a Technical Guidance checklist. In this regard, a protocol that focuses on a checklist of questions to facilitate reviews of DERA/WOE should emphasize key principles rather than details of construction.

An additional issue concerns harmonization of this WOE guidance (and additionally, the remainder of the DERA guidance) with the four volume guidance manual for sediment assessments enacted by Technical Guidance 19. Technical Guidance 19 is a one page document with links to the four volume guidance manual. The four volume guidance manual is somewhat incompatible with DERA and the WOE guidance, both of which consider the evaluation of lines of evidence in sediment assessments. For example, MOE's Technical Guidance 19 adopts the entirety of the United States Environmental Protection Agency 8-step process (which was considered but not adopted directly in other SABCS

guidance), and the terminology used throughout is also different. As coexisting guidance manuals may lead to confusion on the part of risk assessors and Contaminated Sites Approved Professional (CSAP) Roster reviewers, the prioritization of guidance should be clarified.

As discussed above, there are a number of places in the attribute weighting and response magnitude evaluations where decision rules may be informed by policy determinations. Provincial policy is in a state of flux, and WOE details may be influenced by future developments. There is a disclaimer in DERA regarding the role of policy determinations; this would also be applicable to WOE guidance, and practitioners should routinely consult updates to policy determinations.

Finally, MOE will need to consider how to provide regulatory guidance around the use of the narrative descriptors (i.e., negligible, low, moderate, and high). One particular challenge is that “high risk” is also being used in the context of Protocol 12. Providing clarity about the correct application of these narrative descriptors (in a manner that can be evaluated by Roster review) is an important action that will improve the standardization of practice by different risk assessors. It may be appropriate to define “narrative descriptors of magnitude of response” and “narrative descriptors of risk” in the Administrative Guidance and amplify the existing definition of “high risk” to acknowledge its multiple uses.

FUTURE ADDITIONS

The main area of additional guidance development lies in the domain of the linkage to risk management. At the outset, the Task Force recognized the importance of linking the WOE process to the broad site management objectives. Consistent with the philosophy of beginning with the end in mind, it was acknowledged that the “how” of WOE should be influenced by the “why”. The scope of the Exponent deliverable was limited to two of the three steps of the overall risk assessment framework, leaving room for future development of guidance related to risk management.

One issue which could be further clarified in future guidance is the requirement of agreement or consensus between the risk assessor and risk manager. The degree to which risk assessors and risk managers will reach a consensus on WOE elements (such as LOE weightings and magnitude of response categories) during the Problem Formulation review process should be investigated further. In addition, the linkage of these decision-making elements to the role of the Contaminated Sites Approved Professional (CSAP) Process and the current constraints (e.g., requirements for pre-approvals under Protocol 6; degree of Ministry involvement at high risk sites) should be clarified.

CONCLUSIONS

Overall, we believe the guidance does a solid job of balancing the different perspectives of the Task Group. This was a particularly challenging task because there is a tendency for some experienced risk assessment practitioners

to desire the flexibility to customize a WOE framework to reflect site-specific understanding, whereas others desired a clear default procedure to provide consistency. The Appendix IV as written provides a mechanism that meets the needs for a majority of practitioners. Whereas there are differences of opinion regarding the optimal default level of prescription, the Task Force agrees that flexibility is possible within the Exponent Framework provided that there is suitable rationale provided.

ACKNOWLEDGEMENTS

We thank the SABCS for their support of this effort, and for the opportunity to comment on the final Exponent report. We also thank Exponent for their dedication and responsiveness to our feedback. I would also like to personally thank the individual Task Force contributors for their thoughts, patience, open-mindedness, and energy throughout the life of this project.

The following individuals participated in the WOE Task Force:

- Gary Lawrence, Golder Associates (Task Force Chair)
- Paul West (SABCS Project Manager)
- Will Gaherty, Pottinger Gaherty Environmental Consultants (SABCS Board Liaison)
- Beth Power, Azimuth Consulting Group
- Ryan Hill, Azimuth Consulting Group
- Blair McDonald, Golder Associates
- Remi Odense, Ministry of Environment
- Ute Pott, Environment Canada
- Joline Widmeyer, EBA Engineering Consultants
- Doug Bright, AECOM

Exponent®

Guidance for a Weight of Evidence Approach in Conducting Detailed Ecological Risk Assessments (DERAs) in British Columbia

Prepared for

Paul West, Ph.D., FCIC, PChem
Chair, Science Advisory Board for
Contaminated Sites in British Columbia
University of Victoria, Victoria
British Columbia, V8W 3V6

Prepared by

Exponent
15375 SE 30th Place
Suite 250
Bellevue, WA 98007

June 2010

Doc. no. 0906563.000 04F1 0510 MM28

Contents

PREFACE

The purpose of this report is to provide guidance for a weight-of-evidence (WOE) approach for conducting Detailed Ecological Risk Assessments (DERAs) in British Columbia. While there are many definitions of WOE, it is defined here as the process by which measurement endpoints, which are closely linked to lines of evidence (LOEs) are integrated to evaluate the likelihood and magnitude of ecological risk for each assessment endpoint. This guidance was written to help risk assessors develop WOE assessments that are objective, transparent, and scientifically rigorous. The guidance presented herein is the default WOE approach for conducting DERAs in British Columbia; however, an alternative approach may be employed in conjunction with a clear and defensible communication of its structure and merits.

The WOE guidance is provided in three sections and should be inserted directly into the existing DERA guidance document[1]. The first section is guidance on how to use WOE to select measurement endpoints, which is conducted as part of the problem formulation stage of the DERA. Therefore, this section should be placed in the DERA guidance document as Section 3.7.3. During the problem formulation phase of the DERA, the risk assessor should assign weighting factors to each of the LOEs with respect to its quality and its relevance to an assessment endpoint, and then select those with the highest weights for use in the risk analysis. The second section is guidance on how to use WOE in the risk characterization stage. Therefore, this section should replace the existing Section 6.3 in the DERA guidance document. During the risk characterization stage, the weight given to each LOE is revisited and possibly adjusted if unforeseen events during the collection or analysis of a sample have affected the quality or quantity of data, the appropriateness of an established data analysis method, and/or the sensitivity or representativeness of the LOE. Also during this stage, the magnitude of response for each LOE is determined. Taken together, the weight and magnitude of response of each LOE are used by the risk assessor to reach a conclusion regarding risk. The third section is an appendix that provides detailed guidance on the WOE approach and examples at key steps in the WOE process. This section should be inserted in the DERA guidance document as Appendix IV.

As part of developing the WOE approach described herein, Exponent performed a review of selected literature on WOE approaches published between 2000 and 2009. This is provided as an appendix to this report. However, it is not Exponent's intention that this literature review on selected WOE papers be inserted into the DERA guidance document.

---

[1] Golder Associates Ltd. 2008. Detailed Ecological Risk Assessment (DERA) in British Columbia — Technical Guidance, Final 2008 Revision. September 3, 2008.

3.7.3 Using Weight-of-Evidence (WOE) to Select Measurement Endpoints

Each assessment endpoint may have one or more measurement endpoints that can be used to provide information about the assessment endpoint, and multiple methods may be used to derive the same information. For example, assessing risks to a receptor often includes: (1) measuring the concentrations of a COPC in environmental media (soil, sediment, water) to provide information about exposure; and (2) measuring biological or toxicological responses in natural or controlled environments to provide information about effects. Laboratory bioassays evaluate toxic responses of indicator organisms and field surveys provide information about population or community dynamics, both of which can be used to infer the likelihood of risk to single species or a community of organisms. Each of these measurements contributes a different type of understanding to the risk analysis, and not all measurements are of the same quality or relevance. For example, chemical measurements may be made using a screening method (*e.g.*, hand-held XRF to measure soil metals), a general laboratory approach (*e.g.*, inductively coupled plasma spectroscopy), or a highly-specific method (*e.g.*, atomic absorption spectroscopy). Within each suite of tools applied, there are also specific variations in the protocols, target analytes, or endpoints (*e.g.*, mesh size for invertebrate sieving, toxicity test species and duration, choice of chemical groups or congeners for quantitation).

The differences in type, quality, and environmental relevance of candidate measurement endpoints (and their associated "lines of evidence" [LOEs]) have a significant bearing on their applicability to the risk assessment and conclusions. Therefore, following the initial scoping described in Section 2.4.2 and during the problem formulation phase of the DERA, the risk assessor should "weight" each of the LOEs with respect to its quality and its relevance to an assessment endpoint, and then select those with the highest weights for use in the risk analysis. Evaluating the LOEs during problem formulation (whether through formal weighting procedures or qualitative screening) provides the risk assessor an opportunity to look for additional LOEs if the initial list returns only those of relatively low weight. The final selection of LOEs should also account for the selected risk management approach (for example, screening level or small site assessments may need less certain or precise risk analyses than detailed assessments and complex sites).

This section provides a general overview of such a process; detailed guidance is provided in Appendix IV. During the risk characterization (see Section 6.3), the risk assessor must revisit the weightings that were assigned to the LOEs during problem formulation, as they may change depending on whether or not the data collected met the study objectives. Note that the

WOE framework presented herein describes the weighting and integration of LOEs for a single assessment endpoint (*i.e.*, potential receptor of concern), and that the same process must be completed for each assessment endpoint addressed in the ecological risk assessment.

3.7.3.1 Guiding Principles

Guiding principles for WOE assignments (irrespective of the environmental media under investigation) include:

- Whenever possible, LOEs incorporated in the WOE for each assessment endpoint should incorporate multiple, broad types of information, such that the analyses are not limited to the same perspective (*i.e.*, conceptually redundant). For example, it is desirable to include both: (a) laboratory studies conducted under controlled conditions; and (b) field measurements of resident populations (Chapman and Hollert 2005). Also, it is desirable to include both: (a) evaluations of individual chemicals (may be obtained from studies in the literature); and (b) evaluations of chemical mixtures representative of site conditions. These different types of LOEs provide complementary information that strengthens the confidence in reaching robust risk conclusions. Laboratory-based LOEs provide the ability to measure contaminant-related effects under standardized conditions that reduce the influence of other non-contaminant-related stressors. In contrast, field-based LOEs capture information about adverse effects under realistic but potentially variable exposure conditions. Each LOE reflects trade-offs among competing considerations such as: site-specific relevance versus controls for variation; relevance versus cost; and specificity (reductionism) versus ecological realism. These LOEs may be measured simultaneously or used in a sequential (tiered) approach (Hull and Swanson 2006), in which particular LOEs are implemented depending on the results of others (see Section 2.4.1).
- Measurement endpoints must be clearly defined when applying the WOE approach. See Section 3.7.1 for definitions and examples of good measurement endpoints. They should represent some quantifiable effect on an ecological receptor that is clearly related to the assessment endpoint. LOEs are closely related to measurement endpoints, although the definitions differ slightly (see Section 2.1 of Appendix IV). LOEs should be specific to the COPCs, receptors of potential concern, and exposure pathways under investigation.
- The risk assessor should consider LOEs that can be used to evaluate causation (*e.g.*, *in situ* measurements of toxicity, measurements of contaminant body residues as direct evidence of exposure, diagnostic tests such as toxicity identification evaluation [TIE]

procedures). These causality investigations are useful for resolving apparent disagreements among LOEs. They can be considered from the outset or implemented in a tiered fashion. Criteria for evaluating the causality of measured effects can be established (see, for example, Lowell *et al.* 2000[1]). Where causation cannot be directly evaluated using a LOE, multiple alternative hypotheses to explain observed responses should be considered.

- It is necessary to establish an *a priori* framework for integrating different LOEs that support a single assessment endpoint, as not all LOEs have the same degree of specificity and each may measure a slightly different attribute of the ecological receptor of concern. The framework as described in this DERA guidance should include both the magnitude of the response observed in each LOE and its relative weight (*i.e.*, how well it applies to the assessment endpoint and establishes causality, as well as the quality of the data). However, there is no single framework applicable to all sites or assessments; what is presented here may be modified by the risk assessor as needed, with appropriate justification and description provided in the risk assessment report.
- The approach for characterizing magnitude of response of each LOE should be defined and documented during the problem formulation to avoid bias during the risk characterization. The magnitude of response can be characterized on the basis of categorical magnitude (*e.g.*, negligible, low, moderate, or high, as described in Section 6.5), numerical magnitude, spatial scale, level of organization, permanence, probability or frequency of occurrence, *etc.*
- WOE is not an entirely prescriptive methodology, but relies on best professional judgment for selection of the LOEs. However, it is imperative that the risk assessor provide clear and transparent information about why certain LOEs were used and others were not, and how the information was integrated into the final risk characterization for each assessment endpoint. This requirement is included in the WOE framework described herein. Further, the risk assessor may view the overall WOE framework presented herein as a suggested (default) approach, and therefore may consider and apply alternative approaches that are more appropriate for certain sites

[1] Lowell et al. (2000) established *a priori* causal criteria for evaluating different LOEs in a WOE approach for assessing effects of contaminants on northern rivers. Criteria included: spatial and temporal correlation; plausible explanation linking stressor and effect; experimental verification of stressor cause-effect relationship under controlled conditions; strength of correlation; specificity of the effect to the COPC; evidence of COPC exposure in the body of the ROPC; and consistency of association across other studies within the region and in analogous studies in other regions.

and conditions. If this is the case, the reasoning and methods must similarly be provided in order to make the assessment transparent and defensible.

In summary, WOE assessments must be: (a) objective, (b) transparent, and (c) scientifically rigorous. These considerations are in addition to the factors considered during the Risk Management process; specifically that the WOE process be linked to relevant legal and policy considerations, that the assessment consider practical site constraints and needs, and that it dovetail with the conceptual scientific approach applied to the site. Overall, the WOE is designed to answer questions of management relevance and to proceed to the level of certainty needed for site management purposes. Because well-executed WOE approaches are transparent and rely on easily understood weighting factors, they provide a useful communication tool for informing risk managers and other interested parties of the risk assessment results. Use of the framework also makes it possible for technical reviewers of the risk assessment to better follow the logic used by the risk assessor to reach the stated conclusions.

3.7.3.2 Linkage to Risk Management

The WOE approach in British Columbia has been conceptualized in three stages:

- Risk Management (framing the issues and conceptual approach)
- Problem Formulation (selection and evaluation of LOEs, including assignment of weights)
- Risk Characterization (revisiting of weights assigned to the LOEs, synthesis of LOEs, coherence evaluation, and summary of findings).

The WOE approach begins during the scoping of a risk assessment prior to the onset of formal analysis or study design. The practitioner usually works alongside a Risk Manager to define the broad study objectives, output needs, and constraints that would affect the range of approaches applicable to the site. The WOE approach is further developed and applied in the problem formulation stage to select the LOEs for each assessment endpoint and to reach consensus among the risk assessors and risk managers about the required level of confidence in the final risk characterization. The process of assigning weights to each LOE requires an explicit acknowledgment by the risk assessor of the expected quality of the data to be collected and the strength of the association of the LOE with the assessment endpoint, thereby providing a means for determining the strength of the final risk assessment. Because not all risk assessments require the same degree of certainty, selection of LOEs is highly site-specific (*e.g.*, small sites with low-cost clean-up methods do not require complex, highly definitive analyses whereas areas with higher value and greater clean-up costs may need more certainty in locating areas of risk and specific identification of causality). Sites also differ in terms of the volume,

quality, and relevance of site-characterization data available at the beginning of the risk assessment process; the problem formulation stage is used to evaluate existing data and determine how candidate LOEs could be used to fill information gaps. The process of weighting and selecting LOEs provides a transparent record of how the risk assessor considered data quality, causality, uncertainty, and cost/time in selecting the LOEs used in the risk analysis. During the risk characterization stage, weighting factors are re-evaluated to determine if they should be changed as a result of any problems encountered during the data acquisition process, and the weighted LOEs are combined into the final risk assessment (see Section 6.3). As noted, more detailed guidance is provided in Appendix IV.

3.7.3.3 Considerations for Selecting and Weighting LOEs

The selection of LOEs is linked strongly to the articulation of assessment endpoints. The value (and numerical weight) of an individual LOE is in large part a reflection of how closely it corresponds with the assessment endpoint. As such, if the assessment endpoint is clearly defined, the appropriate LOEs will be easier to identify, the rationale for selection will be clearer, and the decision criteria for interpreting responses will be more understandable. If the assessment endpoint is not clearly defined, then the assessment and selection of LOEs will be unclear and may lose value for supporting site management decisions. Guidance for selecting appropriate measurement tools and models to be used as LOEs is provided in Sections 3.7.1 and 3.7.2. In addition, the selection of LOEs is informed by the weights given to each potential LOE. The risk assessor should conduct a thorough analysis of potential LOEs (for each assessment endpoint) prior to weighting them; the final list of selected LOEs is then determined using the weights.

LOEs vary with respect to several attributes, including: (a) strength of association between the LOE and the assessment endpoint; (b) their sensitivity and specificity; (c) the quality of the data and study design; (d) their representativeness; and (e) the degree of correlation observed between the level of stressor(s) and the magnitude of response(s) or effect(s), *i.e.*, causality. Appendix I provides a list of Direct Measurement Tools for conducting DERAs and Appendix II lists some models that may be used to estimate exposure or effects. These Direct Measurement Tools and Models fall into one of three categories:

1. Measures of exposure (*e.g.*, chemical concentrations in abiotic media and in organisms/tissues; food chain models; measures of bioavailability)
2. Direct measures of toxicity (*e.g.*, bioassays or laboratory studies to develop dose-response relationship models)
3. Ecological effects measures (*e.g.*, community or population metrics and models).

The LOEs for a specific risk assessment may be selected from this list as well as other sources (including the scientific literature, other compilations of standard test methods, *etc.*) and are assigned weights based on the above attributes. If the risk assessor decides not to use the Direct Measurement Tools or Modeling Tools in Appendices I and II, the risk assessor must provide the following information for each tool: a description of the tool, to which ecosystem it could be applied, the frequency of its use in DERAs, the benefits of using this tool in a DERA, the common issues with this tool when used in a DERA, and a list of resources for use of this tool. Using this information, the risk assessor should be able to weight the LOE for selection in the same way as those listed in Appendices I and II.

After assigning weights, the LOEs with the highest weights are selected. Additional LOEs may be considered if none of those on the initial lists are highly weighted. If the risk assessor chooses not to select at least one LOE for each LOE category (measure of exposure, direct measure of toxicity, and ecological effect measure), a justification should be provided as to why they have not done so and how the risk characterization will address the lack of information. Appendix IV provides further guidance on how this process can be used to develop a list of LOEs and associated weights for each assessment endpoint. As noted, as long as the reasoning and methods are presented clearly, the risk assessor may employ an approach that differs from the framework presented herein. For example, a risk assessor may decide to apply the LOEs in a tiered fashion. LOEs with lower weights that involve fewer resources or could be done more quickly might be used first, or the risk assessor might establish that a positive result in one LOE would trigger the need to conduct a second to incorporate additional receptors of potential concern. Examples of tiered ecological risk characterizations are described by Fairbrother (2003) and Hull and Swanson (2006) and the guidance for incorporating this approach is discussed further in Section 2.4.1. The risk assessor should include a justification or narrative on the choice of LOEs that includes timelines and/or cost-efficiency issues, if these were a factor in the selection process. At complex sites, the risk assessor may consider using a quantitative method to score and select LOEs, whereas small sites may simply use a conceptual approach. In any case, a narrative explanation of how LOEs were selected must be provided.

References cited

Chapman and Hollert 2005 – already in the DERA reference section

Fairbrother A. 2003. Lines of evidence in ecological risk assessment. Human Ecol Risk Assess 9:1475–1491.

Hull, R.N. and S. Swanson. 2006 -- already in the DERA reference section

Lowell et al. 2000. – already in the DERA reference section

6.3 Using Weight-of-Evidence in Risk Characterization

Risks are characterized with respect to each defined assessment endpoint. A weight-of-evidence (WOE) approach means that the assessment of risks involves considering more than one line of evidence (LOE; closely linked to the measurement endpoint) for a particular assessment endpoint. During problem formulation (Section 3.7.3), each LOE used to characterize risk to an assessment endpoint was assigned a preliminary weight based on five attributes, including: (a) strength of association between the LOE and the assessment endpoint; (b) their sensitivity and specificity; (c) the quality of the data and study design; (d) their representativeness; and (e) the degree of correlation observed between the level of stressor(s) and the magnitude of response(s) or effect(s). However, unforeseen events during the collection or analysis of a sample can affect the quality of data, the appropriateness of an established data analysis method, and/or the sensitivity or representativeness of the LOE. Therefore, for each assessment endpoint addressed in the ecological risk assessment, the weightings for each LOE should be revisited during the risk characterization stage and adjusted to reflect the quality and adequacy of the data collected. This is particularly important for the attribute concerning degree of correlation between stressor and response (*i.e.*, causality), which is described in detail in Appendix IV.

LOEs will vary in the magnitude of their response, and the present WOE approach captures this aspect of the LOEs along with the weights that have been assigned according to their attributes. The magnitude of response can be characterized on the basis of categorical magnitude (*e.g.*, negligible, low, moderate, or high), numerical magnitude, spatial scale, level of organization, permanence, probability or frequency of occurrence, or other metrics. The risk assessor should prepare a narrative that clearly describes how the weights and magnitudes of response for the various LOEs informed their decision about risk. For instance, each LOE may provide a different type of information (*e.g.*, the level of exposure resulting in a single organism response versus a description of current effects [or lack thereof] on the local biodiversity). Therefore, the risk assessor must take into account the different types of information provided by the LOEs, giving consideration to which LOE(s) provide the best information as determined by their weights. This process is described in greater detail in Appendix IV. As noted in the guidance for problem formulation (Section 3.7.3), the detailed approach presented in Appendix IV is intended as a suggested or "default" framework; an alternative approach may be employed in conjunction with a clear and defensible communication of its structure and merits.

Key issues for the DERA practitioner

- How were the final weights associated with each LOE determined?
- What magnitude of risk is associated with each LOE?
- How do the responses of each LOE inform the risk decision, given their relative weights and response magnitudes?

Content for the DERA

- A narrative should be provided to describe the process used to assign final weights to each LOE. A separate narrative may be needed for each assessment endpoint.
- Tabular and/or graphical presentation of the LOE weights and magnitudes of responses is encouraged as a visual communication of the process described.
- Additional narrative should be provided to describe the process used by the risk assessor to combine the information from all the LOEs into a final risk conclusion for each assessment endpoint. This should include information for how the relative weights were taken into account, and should acknowledge that different LOEs for the same Assessment Endpoint address different aspects of potential risk (exposure, direct measures of toxicity [laboratory], ecological [field] effects measures).

As an example of the different types of information that may be available, one LOE ("LOE 1") may represent a comparison of the concentration of a COPC in sediments to a benchmark for benthic invertebrates, LOEs 2 and 3 may be sediment bioassays, and LOEs 4, 5, and 6 may be metrics of benthic community structure. If those LOEs having high weights show a negligible or low response (potentially coupled with low-weighted LOEs showing a higher response), then the risk assessor may conclude that the overall risk to benthos from contaminated sediments (the assessment endpoint) is negligible to low. Attributes such as strength of association to the assessment endpoint and study design, among others, are taken into account in the weighting process (see Appendix IV for details).

The risk assessor should write a narrative explaining the process that was used to develop the final weights for the LOEs and the logic used to combine the various LOEs into a risk conclusion. This is analogous to writing the results and discussion sections of scientific papers and is intended to help other reviewers or risk managers understand how the risk assessor reached their conclusions based on the evidence in hand. The narrative can be used to help

reach agreements, identify disagreements, and identify aspects of the risk assessment that require additional clarity.

As described in Appendix IV, LOEs should be placed into one of the three LOE categories: measure of exposure, direct measure of toxicity, or ecological effect measure. In the narrative, the results of a given LOE should be discussed relative to other LOEs within each one of these categories, and an explanation provided about the kind of information each category has brought to the risk assessment. The LOEs are then considered across categories, to assess whether the ecological effects are related to measures of toxicity and exposure. This puts effects into an ecological context specific to the site and also looks for explicit dose-response relationships. If there is an unequal number of LOEs in the categories, the risk assessor must explain how they will integrate the LOEs such that a balanced approach to the WOE is achieved. Additionally, the risk assessor should consider and acknowledge processes and endpoints for which no formalized LOEs were developed, which were therefore not considered in the WOE procedure. This process involves the use of professional judgment, which is discussed further in the next section (Section 6.4). Uncertainties associated with this lack of information, as well as with the measured LOEs, should be clearly described in the narrative and incorporated into the Uncertainty Assessment (see Section 6.6).

6.3.1 Communicating Risks

Because LOEs provide different kinds of information with varying degrees of confidence, and because LOEs sometimes provide conflicting results, it is the responsibility of the risk assessor to clearly communicate to the risk manager how the information collected led to the conclusion about risk(s). This is done through a two-step process, where the LOEs with their associated weights and magnitude of responses are laid out in an objective manner, generally through the use of a matrix or other graphic. The second step is a coherence analysis to explain the ecological relevance of the various endpoints, how they vary with time and space, and how they collectively inform the risk conclusions. Uncertainties associated with the LOEs, either individually or grouped by type (exposure, laboratory, field) are also described, along with a quantitative or qualitative discussion of the probability of false positive or negatives. Further detail about presentation of the LOEs during risk characterization is in Appendix IV.

Establishing causation and risk is a particular construct within the regulatory and legal framework of risk assessment. In doing so, expert judgment about the plausibility of a particular causal relationship is juxtaposed with the regulatory need for decision-making within the context of societal goals for protection of ecological endpoints. While ecological risk assessments are not conducted under the strict rules of legal evidence, the same principles of allowing the preponderance of evidence to drive decision-making apply (see Jasanoff 1995). Thus, conflicting results from one or more LOEs can be tolerated, as long as the risk assessors

clearly describe why the totality of the evidence points toward a particular risk level and causative agent(s). A case is built by aggregating several types of evidence, none of which may be conclusive on its own, into an effective description of the potential for a COPC to cause harm (its hazard) and the actual *in situ* level of exposure to the chemical(s). Alternative hypotheses may need to be evaluated to build a convincing case that one or more of the COPCs are the most plausible cause for the observed effects. In general, less weight is given to inferences based only on chemical structure (*e.g.*, use of structure-activity relationships) or laboratory studies in the absence of corroborating field data. Laboratory studies provide information on causal mechanisms, which supports inferences about the hazard of COPCs and the potential for detrimental effects to occur. Site-specific evidence of exposure (generally through measurement of COPCs in soil, water, and sediment) is required to show that the potential hazard from a COPC has caused or will cause an effect at a particular site. The degree to which a risk characterization requires explicit mechanistic or causal linkages is a function of the management goals for the site. In some cases, the preponderance of evidence may be sufficient to advance understanding to a management decision point, whereas other contexts require more explicit linkages between cause, effect, and ecological implications.

When presenting the results of an assessment, the risk assessor should strive for transparency, clarity, consistency, and reasonableness (TCCR) (USEPA 2000). Using a chart (such as that shown in Text Box 5 of Appendix IV) to summarize the results and strength (weight) of each LOE can have a tremendous impact on clarity, as it presents a visualization of the results of the analysis. The narrative description of the weighting of the attributes and the magnitude of response for each LOE will allow the technical reviewers to judge the validity, consistency, and reasonableness of the risk conclusions based on clearly described logic. Using this approach will increase the likelihood of acceptance by technical reviewers, risk managers, and other interested parties.

References cited

Jasanoff, S. 1995. Science at the Bar: Law, Science, and Technology in America. Harvard Univ. Press, Cambridge, MA. 285pp.

USEPA. 2000. Risk characterization handbook. EPA 100-B-00-002. U.S. Environmental Protection Agency, Office of Science Policy, Washington, DC. 60pp.

APPENDIX IV

WEIGHT-OF-EVIDENCE APPROACH

TABLE OF CONTENTS

1.0 INTRODUCTION

While there are many definitions of weight-of-evidence (WOE), WOE is defined here as the process by which measurement endpoints, which are closely linked to lines of evidence (LOEs), for a particular assessment endpoint are integrated to evaluate the likelihood and magnitude of ecological risk. Examples of assessment endpoints and measurement endpoints are provided in Section 3.7 of the DERA guidance. A site-specific WOE approach is first developed in the problem formulation stage. Various LOEs are considered and weighted, and the rationale for the selection of specific LOEs as well as the weights assigned to them are fully described and documented. In the risk characterization stage, after data are collected, results for the LOEs are integrated for each assessment endpoint in order to reach conclusions regarding risk. This appendix provides a detailed description of the four-step WOE procedure for use in the DERA. The risk assessor may use an alternative approach depending upon site-specific factors; in this case, the reasoning and methods must be communicated in the final narrative to complete a transparent and defensible WOE assessment.

The following steps are conducted separately for each assessment endpoint. They are briefly presented below; more detail is given in the following sections. An alternative WOE approach may be used provided the details of the procedure are clearly explained (*e.g.*, in terms of the specific attributes used to evaluate relative merits of LOEs, degree of quantification of these attributes, formality of decision rules for ranking or scoring of attributes). However, the broad conceptual process for conducting WOE following the four step process (outlined below) should be universally applicable to all risk assessments.

1. **Select, evaluate, and assign weights to each LOE—**LOEs may vary in several respects, such as the extent to which they relate to the assessment endpoint, their sensitivity and specificity, the quality of the supporting data and study design, their representativeness, and the degree of correlation observed between the level of stressor(s) and the magnitude of response(s), (*i.e.*, causality). In this step, potential LOEs are evaluated, screened for relevance to the ERA, and may be weighted based

on the above attributes. Guidance for selecting appropriate measurement tools and models to be used as LOEs is provided in Sections 3.7.1 and 3.7.2 of the DERA Guidance Manual; Appendix I provides examples of Direct Measurement Tools for conducting DERAs and Appendix II has examples of models that may be used to estimate exposure or describe responses. During the problem formulation stage, potential LOEs that are applicable to each assessment endpoint are selected from these lists or from other sources. Weights are assigned to the various LOEs and the rationale for the selection and weight of each LOE is documented. Documentation is also required if certain types of LOEs are not selected for use in the risk assessment. This process provides a transparent record of how the risk assessor considered LOEs and their attributes when selecting the LOEs for each assessment endpoint. The greater the degree of *a priori* evaluation of potential LOEs, the greater the confidence that the interpretations of the LOEs in subsequent steps are not arbitrary.

2. **Adjust weights of each LOE if appropriate—**After data are collected, weights assigned to certain attributes of an LOE may change. This reflects the practical implications of investigation findings, which may require adjustments from *a priori* evaluations that were based on an idealized study design. For example, weights may be lowered if data quality objectives could not be met, or if sampling efforts did not yield the desired number and type of samples. During this step, weights assigned to each of the LOEs should be re-evaluated and adjusted as necessary, with a clear justification provided. This step is conducted early in the risk characterization stage, and should be conducted prior to the detailed analysis of endpoint data, rather than used as a means of adjusting results.

3. **Determine the magnitude and type of the response/effect of each LOE—** Following a review of the collected data, narrative descriptors are used to describe the magnitude of response (*e.g.*, negligible, low, moderate, or high). For some alternative WOE applications, it may also be necessary to formally evaluate some other aspect of the response/effect data, such as evidence for causality or level of uncertainty. This step is conducted in the risk characterization stage.

4. **Integrate LOEs based on their magnitudes of responses, relative relevance, and coherence—**During the final step of the risk characterization stage, the LOEs for an assessment endpoint are synthesized using a transparent and logical method of combining results. The default method entails placement of LOEs on a matrix that graphically displays the weight and magnitude of response of each LOE. The graphical display can be used to evaluate concurrence among the LOEs. This step provides and displays the logic that serves as the basis for a narrative description of risk to the associated assessment endpoint. With respect to contaminated sites, this logic focuses on the critical question, "Are chemicals and associated biological/toxicological responses resulting in risks to valued ecological resources?" The risk assessor should consider how the LOEs inform this question. For example, it is possible that one LOE shows a strong chemical-related response, but other LOEs indicate that the response is not related to chemicals and associated toxicity. The logic behind the various combinations of LOEs used in the benthic triad (*e.g.*, Chapman and Anderson 2005) is an example of this aspect of LOE integration. The narrative should state what information is provided by each LOE and how that information is considered with respect to judging whether site-related chemicals are posing risks to receptors of concern.

2.0 DETAILED WOE PROCEDURE

This section provides step-by-step, detailed guidance and illustrative examples for conducting WOE in DERAs.

2.1 Problem Formulation

During problem formulation, LOEs that will be used to evaluate the likelihood and magnitude of ecological risk for each assessment endpoint are selected. LOEs must be specified in detail so that the quantifiable property to be measured is apparent and can be evaluated for its relevance to the contaminant(s) of potential concern (COPCs), receptor(s) of potential concern (ROPCs), and exposure pathway(s). LOEs should represent some quantifiable effect on an ecological receptor from a measurable occurrence of a stressor, as defined by the assessment endpoint. A line of evidence

differs from a measurement endpoint in that the latter may be applied to multiple assessment endpoints while each LOE is applicable to only one assessment endpoint and risk hypothesis. In other words, once a measurement is used to assess a particular assessment endpoint, it becomes an LOE for the specified assessment endpoint.

The selection of Direct Measurement Tools (Appendix I) or data required for modeling (Appendix II) and selection of LOEs are inextricably linked. As more scientific measurement tools become available, and as our mechanistic understanding of the processes underlying these tools improves, the list of possible LOEs grows and LOEs with higher weights can be identified. Several potential LOEs should be considered for each assessment endpoint. In fact, a core objective of Step 1 is to provide confidence to readers of the DERA that selection of LOEs was not arbitrary and that an objective evaluation of candidate tools was applied in screening LOEs. If the risk assessor decides to use tools that are not listed in the DERA appendices, the following information must be provided for each tool: a description of the tool, to which ecosystem it could be applied, the frequency of its use in DERAs, the benefits of using this tool in a DERA, the common issues with this tool when used in a DERA, and a list of resources for use of this tool. Using this information, the risk assessor should be able evaluate the LOE for selection in the same way as those listed in Appendices I and II.

The Direct Measurement Tools and models fall into one of three categories:

1. Measures of exposure or bioavailability (*e.g.*, chemical concentrations in abiotic media and in organisms/tissues; food chain models; binding/partitioning models)
2. Direct measures of toxicity (*e.g.*, bioassays or laboratory studies to develop dose/concentration response relationship models)
3. Ecological effects measures (*e.g.*, community or population metrics and models).

The risk assessor should attempt to select LOEs from each category for a given assessment endpoint. If the risk assessor chooses not to select at least one LOE for each LOE category (measure of exposure, direct measure of toxicity, and ecological effect measure), they must provide justification as to why they have not done so and how the

WOE approach will address the lack of information. It is not mandatory to apply an LOE from each category to support a defensible DERA; however, the implications of excluding any category must be considered, particularly in terms of describing the resulting uncertainties in the risk characterization.

2.1.1 Select and Assign Preliminary Weights to Each Potential LOE (Step 1)

Once candidate LOEs have been identified (using the screening procedure described above) it is necessary provide a preliminary assessment of the relative merits of each candidate LOE in terms of providing meaningful information for the evaluation of the assessment endpoint. To provide an objective process, this evaluation should be clear, logical, and ideally be conducted prior to the formal data analysis. For example, assigning preliminary weights to the LOEs, as described below, will provide a transparent evaluation of the LOEs and aid in the selection of those most likely to provide useful information for assessing risks.

Text Box 1: LOE Examples

Assessment Endpoint = Maintenance of Benthic Community Structure as a Prey Base for the Aquatic Food Web

- Sediment bulk chemistry
- Sediment pore water chemistry
- *Hyalella azteca* bioassay toxicity
- Species sensitivity distribution from literature
- Benthic community survey

Assessment Endpoint = Sustainability of Local Populations of Upland Wildlife

- Soil chemistry
- Plant COPC concentration
- Soil invertebrate COPC concentration
- Small mammal density

As noted above, several potential LOEs should be considered for each assessment endpoint so that the LOEs with the highest weights can be selected. Examples of LOEs for two hypothetical assessment endpoints are provided in Text Box 1. Weights are assigned to each LOE based on a set of attributes that are designed to evaluate the degree of confidence placed in each LOE. In this example, the weights for the LOEs have been assigned a formal classification system and range from 1 to 5 (see Table 1). LOEs with the highest confidence for the most attributes yield the highest weights (*e.g.*, 4 or 5). If an LOE has a low weight (*e.g.*, 1 or 2), the risk assessor should consider whether to use that LOE as part of the risk assessment. Preliminary weights are assigned in the problem formulation stage to assist in the selection of LOEs and are reevaluated in the risk characterization stage after all data are available.

The following attributes and weighting scheme are provided to illustrate the approach, and may be used as an accepted default method for weighting LOEs. If the risk assessor chooses to use other attributes or weighting factors, a narrative must be included in the Problem Formulation section of the risk assessment describing and justifying the alternative approach. In the default system, five attributes (a through e) are considered for each LOE, all of which relate to the relevance of the LOE to inform the assessment endpoint. When LOE weighting is performed, the first attribute, *Strength of Association*, is double counted (by entering it twice into Table 2); this acknowledges the increased importance of this attribute[1]. The weight derived for an LOE is the sum of all six respective attribute scores (two for the *Strength of Association* attribute and one for each of the other four attributes) divided by 6. An example is presented in Text Box 2.

The five specific attributes evaluated in the default LOE evaluation procedure are:

a) Strength of Association between the LOE and the Assessment Endpoint—This attribute refers to the extent to which the LOE is related to or is associated with the

[1] The importance of the linkage between measurement and assessment endpoints is emphasized in most ERA guidance documents including provincial DERA and USEPA documents.

assessment endpoint. LOEs that are directly related to the assessment endpoint should be given a higher weight; LOEs that are indirectly related should be given a lower weight. For example, sediment toxicity observed in a laboratory test that is not directly related to conditions of the populations in the field will be given a lower weight than a detailed field survey of resident benthic invertebrate communities. Generally, LOEs that are based on uncertain laboratory-to-field extrapolations should be assigned a lower weight than field-based LOEs with clear associations with the assessment endpoint. Because this attribute is considered to be more important than the other four, the score for this attribute is counted twice (*i.e.*, double weighted) as shown in Table 2.

Table 1. Example — Definitions of scores applied to LOE attributes in WOE

LOE Attribute	Description of Attribute	Decision Rules for Attribute Scores Used to Weight LOEs				
		1	2	3	4	5
a. Strength of Association	Site-specificity and relevance of LOE to assessment endpoint; linkage based on known biological processes; similarity of effect, mechanism of action, target organ, and level of ecological organization	Biological processes link the LOE to the assessment endpoint only indirectly, yielding a weak association between the assessment endpoint and LOE	Biological processes directly link the LOE to the assessment endpoint and LOE, but the specific effect, target organ, and mechanism of action evaluated are not the same	LOE and assessment endpoint are directly linked and the adverse effect, target organ, and mechanism of action are the same for LOE and assessment endpoint, but the levels of ecological organization differ	LOE and assessment endpoint are directly linked and the adverse effect, target organ, and mechanism of action are the same for LOE and assessment endpoint, and the levels of ecological organization are the same	Assessment endpoint is directly measured and is equivalent to the LOE
b. Sensitivity and Specificity	The degree to which the LOE can detect change above baseline or reference conditions; the degree to which the LOE is specific to certain stressors; the potential for confounding factors to affect interpretation	LOE can detect only extreme responses, and certainty in observed responses is low; Only one or two of the following factors is derived from or reflects the site: data, media, species, environmental conditions, benchmark, habitat type	LOE can detect large changes, but with a low degree of confidence; Three of the six factors listed to left are derived from or reflect the site	LOE can detect large changes with a high degree of confidence, and moderate changes with a low degree of confidence; Four of the six factors listed to left are derived from or reflect the site	LOE can detect moderate changes with a high degree of confidence, and small changes with a low degree of confidence; Five of the six factors listed to left are derived from or reflect the site	LOE can detect small changes with a high degree of confidence; All six factors listed to left are derived from or reflect the site (*i.e.,* both data and benchmark reflect site conditions)
c. Data Quality and Study Design	Extent to which data quality objectives (DQOs) are met; quality of data; use of standard methods	Three or more DQOs are not met OR DQOs barely meet the needs of the risk assessment OR There is no documentation of the reason for not meeting DQO and the impact on the assessment; Method has never been published AND methodology is not an impact assessment, field survey, toxicity test, benchmark approach, toxicity quotient, or tissue residue analysis	Two or more DQOs are not met AND DQOs satisfy the needs of the risk assessment AND Reason for not meeting DQOs and the impact on the assessment are documented satisfactorily; Method is one of six listed methodologies, but the particular application is neither published nor standardized	One DQO is not met AND DQOs satisfy the needs of the risk assessment AND Reason for not meeting DQOs and the impact on the assessment are clearly stated; A standard method exists, but its suitability for this purpose is questionable, and it must be modified to be applicable to site-specific conditions	One DQO is not met AND DQOs are rigorous and comprehensive AND Reason for not meeting DQOs and the impact on the assessment are clearly stated; A standard method exists and it is directly applicable to the LOE, but it was not developed precisely for this purpose and requires slight modification OR the methodology is used in two peer-reviewed studies	All DQOs are met AND DQOs are rigorous and comprehensive; A standard method exists and is directly applicable to the LOE and it was developed precisely for this purpose and requires no modification OR the methodology is used in three or more peer-reviewed studies

Table 1. (cont.)

LOE Attribute	Description of Attribute	Decision Rules for Attribute Scores Used to Weight LOEs				
		1	2	3	4	5
d. Representative-ness	Spatial and temporal overlap among measurements or samples, stressors, and ecological receptors	The locations of two of the following factors overlap spatially only to a limited extent: (1) study area, (2) sampling / measurement site, (3) stressors, (4) receptors, and (5) points of potential exposure; Measurements are collected during a season different from when effects would be expected to be most clearly manifested, AND a single sampling or measurement event is conducted, AND high variability in that parameter is expected over time	The locations of two of the following factors overlap spatially: (1) study area, (2) sampling / measurement site, (3) stressors, (4) receptors, and (5) points of potential exposure; Measurements are collected during a season different from when effects would be expected to be most clearly manifested, OR a single sampling or measurement event is conducted, AND high variability in that parameter is expected over time	The locations of three of the following factors overlap spatially: (1) study area, (2) sampling / measurement site, (3) stressors, (4) receptors, and (5) points of potential exposure; Measurements are collected during the same period that effects would be expected to be most clearly manifested, AND a single sampling or measurement event is conducted, AND moderate variability in that parameter is expected over time	The locations of four of the following factors overlap spatially: (1) study area, (2) sampling / measurement site, (3) stressors, (4) receptors, and (5) points of potential exposure; Measurements are collected during the same period that effects would be expected to be most clearly manifested, AND two sampling or measurement events are conducted, AND moderate variability in that parameter is expected over time	The locations of five of the following factors overlap spatially: (1) study area, (2) sampling / measurement site, (3) stressors, (4) receptors, and (5) points of potential exposure; Measurements are collected during the same period that effects would be expected to be most clearly manifested, AND EITHER two sampling events are conducted and variability is low OR multiple sampling events are conducted and variability is moderate to high over time
e. Correlation / Causation / Consistency	Ability of LOE to demonstrate effects from exposure to stressor and to correlate effects with degree of exposure	LOE response to stressor has not been demonstrated in previous studies, but is expected to be based on demonstrated response to similar stressors; mechanistic linkage absent	LOE response to stressor has been suggested in previous studies, but has not been definitely proven; mechanistic linkage absent	LOE response to stressor has been demonstrated in previous studies, but response is not correlated with magnitude of exposure; mechanistic linkage equivocal	LOE response is quantitatively correlated with magnitude of exposure, but correlation is not statistically significant (or data are insufficient to test for statistical relationships); mechanistic linkage inferred, but not definitive	Statistically significant correlation is demonstrated; clear mechanistic linkage from exposure to response

Source: Adapted from Menzie *et al.* (1996).

Table 2. Example - Assigning weights to each LOE

LOE Attribute	Factors to Consider in Ranking	Attribute Scores (check one box in each row)					
		1	2	3	4	5	Rationale
a. Strength of Association	Site-specificity and relevance of LOE to assessment endpoint; linkage based on known biological processes; similarity of effect, mechanism of action, target organ, and level of ecological organization						
a. Strength of Association	**Note:** The scores for this attribute are entered twice to double-weight this attribute because of its importance						
b. Sensitivity and Specificity	The degree to which the LOE can detect change above baseline or reference conditions; the degree to which the LOE is specific to certain stressors; the potential for confounding factors to affect interpretation						
c. Data Quality and Study Design	Extent to which data quality objectives are met; quality of data; use of standard methods						
d. Representativeness	Spatial and temporal overlap among measurements or samples, stressors, and ecological receptors						
e. Correlation/Causation/ Consistency	Ability of LOE to demonstrate effects from exposure to stressor and to correlate effects with degree of exposure						
Average LOE rank:	Total of scores for the five attributes = (enter total) Average LOE weight = (Circle one) **1 2 3 4 5**						

Notes: There should be one table like this for each LOE.
Provide rationale for selected weight in the corresponding weight box.

b) Sensitivity and Specificity—This attribute refers to the extent to which the LOE is sensitive to the stressor and specific to site conditions. Sensitivity refers to the ability of the LOE to detect a change in the response above natural or analytical variability and uncertainty. Sensitivity is a function of both the accuracy and the repeatability of the test endpoint responses. LOEs that are sensitive measures of a response (*e.g.*, can reliably detect small changes relative to baseline or reference conditions) should be given a higher weight; LOEs that are insensitive measures of a response (*e.g.*, can detect only large changes relative to baseline or reference conditions) should be given a lower weight. For example, if the survival endpoint of the amphipod toxicity test (Text Box 2) routinely yields a minimum significant difference of less than 20% relative to reference sediment, this LOE would be considered to be highly sensitive[2].

Text Box 2: Assigning Weights to Lines of Evidence

Assessment Endpoint = Protection of the Sediment Benthic Community Function

Lines of Evidence	Attribute						Average Weight (divide by 6)
	A Association (entered twice)		B Sensitivity/ Specificity	C Quality/ Design	D Represent -ativeness	E Causality	
1. Sediment chemistry (relative to guideline)	2	2	2	3	3	2	2
2. Amphipod bioassay (relative to reference site)	2	2	4	5	3	2	3
3. Mussel tissue chemistry (relative to tissue residue benchmark)	4	4	3	5	4	4	4
4 Benthic community analysis (using reference envelope)	5	5	3	3	4	3	4

Note: See Table 1 for definitions of scores for LOE attributes.

[2] The minimum significant difference (MSD) is the lowest distinguishable difference that is statistically meaningful. Different toxicity test endpoints have different MSDs, although the MSD of relevance to a study design will depend on the laboratory and test protocol applied. The 20% threshold has been applied in many ecological risk assessment applications, including BC provincial policy for toxicity test interpretation and federal guidance for sediment management (Chapman and Anderson 2005).

Specificity refers to the extent to which data, media, species, environmental conditions, and habitat types used in the study design reflect the site of interest. LOEs that are specific to the site (*e.g.*, biological and chemical data, as well as benchmarks that reflect site conditions) should be given a higher weight; LOEs that are not specific to the site (*e.g.*, biological or chemical benchmarks used to assess data are regional and are not specific to the site) are given a lower weight.

c) Data Quality and Study Design—This attribute refers to the degree to which data quality objectives and other recognized characteristics of high quality studies are met. During problem formulation, it is assumed that the data quality will be adequate and that sampling will be conducted exactly as contemplated in the sampling and analysis plan; however, this attribute is re-evaluated later during the risk characterization following a review of the amount, type, and quality of data generated. LOEs that use precise and standard methods with accepted quality assurance and quality control (QA/QC) procedures should be assigned a higher weight; LOEs that use novel methods, unvalidated data, or imprecise data with questionable QA/QC should be assigned a lower weight. For the purpose of WOE, it is assumed that data considered to be unacceptable (*i.e.*, rejected as erroneous) will be removed from the assessment. For example, a chemical measurement was made using an incorrect method, such as improper solvent, temperature, or run time on the gas chromatograph, rendering the reported values invalid and those data subsequently being deleted from the project database.

d) Representativeness—This attribute refers to the spatial and temporal alignment of the measurements or samples with the stressors and ecological receptors. Stressors, which can vary spatially and temporally, need to be aligned with the biological responses that are measured by LOEs to demonstrate that a measured biological response is related to the stressor and not to some other factor. Additionally, LOEs that capture or integrate natural spatial or temporal variation should be assigned a higher weight (*e.g.*, seasonal benthic community samples collected from areas that collectively represent a concentration or exposure gradient). LOEs that provide transient or sporadic measures of parameters with high spatial and temporal

variability should be assigned a lower weight (*e.g.*, one seasonal measurement of phytoplankton abundance from a thermally-stratified lake).

e) **Correlation/Causation/Consistency—**This attribute refers to the degree to which a significant and reliable statistical association is observed between the level of stressor and the magnitude of response or effect. Whereas statistical correlations provide an indication of potentially significant linkages, such correlations are more compelling when consistent with knowledge of underlying mechanistic factors and/or when correlations are validated through repeated measurements. LOEs that demonstrate that the observed effect is consistently associated with or caused by the stressors should be assigned a higher weight (*e.g.*, the ratio of simultaneously extracted metals and acid volatile sulfides in sediment strongly correlates with amphipod survival in laboratory bioassays with site sediment). LOEs that do not show consistent relationships between stressor and effects and/or where mechanism of action is unknown should be assigned a lower weight (*e.g.*, as shown in Text Box 2, bulk concentrations of cadmium in sediment show no relationship with amphipod survival in laboratory bioassays with site sediment, so "sediment chemistry" (for cadmium) is given less weight). This attribute is considered in greater detail during the risk characterization than in the problem formulation. In evaluating this attribute, practitioners may benefit from consideration of the Hill (1965) summary of considerations for causation; these include strength, consistency, specificity, temporality, presence of biological gradient, plausibility, coherence, experimental evidence, and analogy from similar contexts.

2.1.2 Final Selection of LOEs

As part of the LOE selection process, the risk assessor must indicate how endpoints have been weighted relative to each other. Where a matrix-based scoring evaluation is applied, the evaluation will include the score (1 through 5) assigned to each attribute based on Table 1 and, importantly, the rationale for the selected weight (Table 2). Rationales are particularly important when a less quantitative procedure is applied for evaluation of attributes or LOEs, such that the WOE approach is sufficiently transparent for reviewers.

In the example (Table 2), the numeric scores of the attributes are averaged for each LOE. Attribute "a" (*Strength of Association*) is counted twice in Table 2 because of its importance in weighting an LOE. Therefore, there are six values, which are summed and then divided by six to produce an average. The average value of the attributes represents the overall weight of the LOE. This value is rounded to a maximum of one decimal place.

To determine the final list of LOEs to be used in the DERA, the risk assessor should review the potential measurement endpoints and select those that are deemed to have sufficient weight to be useful in the risk assessment. In some instances, it may be appropriate to sequence the application of LOE such that a subset of LOEs that provide high value of information per unit cost are applied first, followed by tiering of remaining endpoints to refine residual uncertainty. In any case, the risk assessor should clearly state their rationale for selecting, rejecting, or deferring a particular LOE. In practice, overall weights at or below 2 are sufficient to reject an LOE. The list of Direct Measurement Tools in Appendix I, or data required for the models in Appendix II, can be revisited and additional potential LOEs can be weighted to determine the final list of measurement endpoints that will be used in the DERA.

2.1.3 Determination of Magnitude of Response

The way the magnitude of response of each selected LOE will be reported, and the potential ranges that could be expected for that type of measurement, should be defined and documented during the problem formulation to avoid bias during the risk characterization. The magnitude of response can be characterized on the basis of categorical magnitude (*e.g.*, negligible, low, moderate, or high), numerical magnitude (*e.g.*, proportion of population affected, degree of impairment relative to background), or in other ways, as described further below (Section 2.2.2).

2.2 Risk Characterization

During the problem formulation, weights are assigned to each LOE assuming that the studies would produce the best data possible. However, unforeseen events during the

collection or analysis of a sample can affect the quality of data, the sensitivity, or representativeness of the LOE. This section describes how the weights for each LOE may be revised to reflect the quality, quantity, and spatial/temporal distribution of data that were collected. Additionally, the magnitude of the response is evaluated together with the degree of confidence in each LOE to make the final risk determination.

2.2.1 Adjust Weights to Each LOE (Step 2)

Re-evaluating the LOEs in terms of Attributes "b," "c," and "d" (*Sensitivity and Specificity*, *Data Quality and Study Design*, and *Representativeness*, respectively) will result only in keeping the weights the same or lowering them; it is not possible to raise the weight given to the LOE in Step 2 because it was assumed during the problem formulation that the studies would yield the best possible data quality. Examples are provided in Text Box 3.

Text Box 3: Examples of Revised Weights for Lines of Evidence

- Samples could not be collected using the specified methods:
  - Lower the weight given to the *Data Quality and Study Design* attribute; and
  - Consider lowering weight for *Representativeness* depending on the degree to which synoptic measurement was affected by sampling limitations.
- Insufficient tissue mass was obtained so the analytical laboratory had to raise the detection limits:
  - Lower the weight given to the LOE for the *Sensitivity and Specificity* attribute.
- Some samples were lost during field sampling and the remaining samples do not adequately represent the chemical gradient present at the site:
  - Lower the weight given to the LOE for the *Representativeness* attribute.

Attribute "e," *Correlation/Causation/Consistency*, is evaluated in detail in the risk characterization when site-specific data are available. LOEs that demonstrate that the observed effect is consistently associated with or caused by the chemical stressors

associated with a contaminated site should be given a higher weight. LOEs that do not show consistent relationships between chemical stressors and effects should be given a lower weight. In some cases, observations made during sampling (*e.g.*, documentation of highly variable habitat, evidence of physical disruption of soils/sediments, observation of confounding factors such as woody debris) may result in a need to re-evaluate the initial weighting assignments for the WOE. If the combination of LOEs indicates that non-chemical stressors are causing the observed effects, these confounding factors should be discussed. In order to address the risks associated with site-specific chemicals, the risk assessor needs to consider how to best separate those risks from other factors that may be present and unrelated to chemicals at the site. Simple regression and multiple regression analyses may provide some insight regarding the relationship between stressor and effect. In addition, multivariate analysis (Section 6.2 of the DERA Guidance Manual) may be able to detect patterns that emerge from interactions of multiple abiotic and biotic parameters.

2.2.2 Determine the Magnitude of the Response of Each LOE (Step 3)

The magnitude of response of each LOE must also be incorporated into the WOE procedure, as this is an important factor in risk characterization that will serve to guide management actions associated with the type, likelihood, and magnitude of risk. The magnitude of response can be characterized on the basis of categorical magnitude, numerical magnitude, spatial scale, level of ecological organization (individual, population, or community), permanence, probability or frequency of occurrence, etc. The magnitude of response is often scaled to a threshold response considered representative of a background or acceptable conditions (*e.g.*, reference normalization, acceptable effect size, reference envelope comparison). Therefore, the magnitude of response determination may require an evaluation of multiple attributes and is not simply an assessment of the absolute value of the response measure.

The basis for characterizing and defining the magnitude of response (*e.g.*, negligible, low, moderate, or high) for each LOE should be developed and documented during the problem formulation. For example, results of laboratory bioassays may be categorized as

negligible if performance of organisms exposed to site samples is not significantly different from reference site samples, or as high if mortality of organisms exposed to site samples is substantial or is statistically greater than reference sites. In this example, the criteria are based on magnitude and/or statistical significance of the difference between the current site conditions and one or more reference areas, a historical reference condition, or numerical environmental quality guidelines (*e.g.*, published benchmarks for sediment, water, or tissue). Refer to Text Box 4 for an example of how magnitudes of response have been established by others.

Text Box 4: Determining Magnitude of Response

McDonald *et al.* (2007) developed a WOE framework to evaluate potential effects on the aquatic ecosystem of Wabamun Lake (Alberta, Canada) associated with the release of Bunker ''C'' oil after a train derailment. The WOE framework integrated the findings of many LOEs. In the supplemental material, those authors provided criteria for determining the magnitude of risk for each LOE. These criteria were developed *a priori*, and were based on magnitude or statistical significance of the difference between the current conditions and one or more reference areas, a historical reference condition, or numerical environmental quality guidelines (*e.g.*, sediment, water, or tissue guidelines). The table below provides a representative subset of the LOEs and their magnitude of response categories.

Assessment Endpoint	LOE	Benchmark	Magnitude of Response Categories			
			Negligible	Marginal	Moderate	High
Protection of Benthic Invertebrate Community	Sediment chemistry	ISQG, PEL	<ISQG	NA	ISQG-PEL	>PEL
	C. tentans survival	Reference area	<20%	>20% NS	20-50%	>50%
	C. tentans growth	Reference area	<20%	>20% NS	20-50%	>50%
	H. azteca survival	Reference area	<20%	>20% NS	20-50%	>50%
	H. azteca growth	Reference area	<20%	>20% NS	20-50%	>50%
	L. variegatus survival	Reference area	<20%	>20% NS	20-50%	>50%
	Benthos abundance	Historical data	<20%	NA	20-50%	>50%
	Benthos richness	Historical data	<20%	NA	20-50%	>50%
	Benthos evenness	Historical data	<20%	NA	20-50%	>50%
	Benthos Bray-Curtis Distance	Historical data	<20%	NA	20-50%	>50%
Protection of Fish Community	Surface water chemistry	Water quality guideline	<WQG	NA	NA	>WQG
	Fish bile chemistry	Reference area	NS	NA	NA	S
	Fish tissue chemistry	Tissue residue guideline	<TRG	NA	NA	>TRG
	SPMD Chemistry	Reference area	NS	NA	NA	S
	P. promelas survival	Reference area	<20%	>20% NS	20-50%	>50%
	P. promelas growth	Reference area	<20%	>20% NS	20-50%	>50%
	Whitefish hatchability	Reference area	<20%	>20% NS	20-50%	>50%
	Whitefish larval growth	Reference area	<20%	>20% NS	20-50%	>50%
	Whitefish skeletal deformity	Reference area	<50%	NA	50%A	>50%M/S
	Whitefish finfold deformity	Reference area	<50%	NA	50%A	>50%M/S
	Whitefish craniofacial deformity	Reference area	<50%	NA	50%A	>50%M/S
	Whitefish edema	Reference area	<50%	NA	50%A	>50%M/S

From: McDonald *et al.* 2007

Notes:
A = any level of deformity
ISQG = interim sediment quality guideline
M/S = moderate or severe deformity
NA = not applicable
NS = not statistically significant
PEL = probable effects level
TRG = tissue residue guidelines
S = statistically significant
SD = standard deviation
WQG = water quality guideline

2.2.3 Integrate Weight and Magnitude of All LOEs (Step 4)

For each assessment endpoint, corresponding LOEs are synthesized through examination of the relative measures of response (*i.e.*, magnitude) and the quality of information carried by each LOE (*i.e.*, weights). In this manner, the overall strength of response can be compared against the overall reliability of the response measure, and the concurrence of the relationship can be evaluated for multiple LOEs.

LOEs are placed on a matrix that presents the weight along the x-axis (abscissa) and the magnitude of response on the y-axis (ordinate) (Figure 1). For each LOE, the x-axis represents the average weight of the LOE, categorized as 1, 2, 3, 4, or 5, as determined in Steps 1 and 2 (Sections 2.1.1 and 2.2.1). The y-axis represents the magnitude of the response of the LOE determined in Step 3 (Section 2.2.2), which may be categorized as negligible, low, moderate, or high (or some other approach for defining magnitude).

Figure 1. Example WOE Matrix

Assessment Endpoint = _____

		WEIGHT				
		1	**2**	**3**	**4**	**5**
MAGNITUDE OF RESPONSE	**High**					
	Moderate					
	Low					
	Negligible					

Notes: Place a symbol or letter for each LOE on the matrix that integrates the response magnitude and weight.
There should be one matrix per assessment endpoint.

In addition to a graphical representation of the results, the risk assessor may view the matrix as a plane that illustrates convergence or divergence among the LOEs (Text Box 5). The configuration of the LOEs on the plane provides a visualization of the WOE that is used by the risk assessor to reach conclusions regarding risk (or lack of risk). In this example, the sediment concentrations (LOE 1) and one of the bioassays (LOE 2) showed moderate to high responses, although their strength of association to the assessment endpoint (maintenance of benthic community structure as a prey base for the aquatic food web) was less than that of LOE 4. LOE 3, concentrations of COPCs in mussel tissue, was a relatively high weight because it had a high strength of association to bioavailable contaminants and also provided information on potential risks to consumers.

Text Box 5: Example of a Weight of Evidence Matrix

Assessment Endpoint = Maintenance of Benthic Community Structure as a Prey Base for the Aquatic Food Web

		WEIGHT				
		1	**2**	**3**	**4**	**5**
Magnitude of Response	**High**			LOE 2		
	Moderate		LOE 1		LOE 3	
	Low			LOE 4		
	Negligible					

LOE 4 = LOE number

Box color indicates LOE type (orange: measure of exposure; green: direct measure of toxicity, blue: ecological effect measure)

LOE 1 = sediment chemistry relative to guideline (exposure)
LOE 2 = amphipod growth/mortality relative to reference site (laboratory)
LOE 3 = mussel tissue chemistry relative to tissue residue benchmark (exposure)
LOE 4 = benthic community richness (number of taxa relative to reference envelope) (field)

Benthic community richness (LOE 4) was given a slightly lower weight because it can be affected by multiple stressors (not just the COPCs) and was measured only once, so temporal variance cannot be taken into account.

The risk assessor must write a narrative that clearly shows how the LOEs, given their various weights and corresponding magnitudes (effect sizes, scales, probabilities, *etc.*), informed the decision about the conclusion of risk. Text Box 6 provides some examples of the type of language that could be used in such a narrative, but is illustrative and not comprehensive. A full narrative is analogous to writing the results and discussion sections of a scientific paper and is intended to help other reviewers or risk managers understand how the risk assessor reached their conclusions based on the evidence in hand. The narrative can be used to help reach agreements, identify disagreements, and identify aspects of the risk assessment that require additional clarity.

Simply referring to the WOE matrix (Text Box 5; Figure 1) does not provide sufficient information about how the overall conclusion of risk was made. Because each LOE provides a different type of information (*e.g.*, the level of exposure resulting in a single organism response versus a description of the local biodiversity), the risk assessor must take into account the different types of information provided by the various LOEs, giving consideration to the LOEs for which the confidence in the information is greatest, as determined by their weights.

Any other WOE approach, either published or novel, can be used provided it is explained in sufficient detail for a reader to be able to reproduce it with high precision.

Text Box 6: Examples of Language for Use in a Weight-of-Evidence Risk Narrative

Conclusion: There is moderate risk of impairment to the maintenance of the benthic community structure.

Rationale: Given that one LOE showed a low response, one a high response, and two moderate responses, and given that the two LOEs with greatest weights both showed a moderate to high response, the overall risk appears to be moderate.

Two lines of evidence were evaluated to investigate exposure potential (from sediment) and actual (from tissue concentrations). Direct effects (toxicity) to organisms were measured in one LOE and field responses were measured in another LOE.

Exposure:

Chemical concentrations (LOE 1) in sediment are moderate to high relative to sediment quality guidelines. There are no physical/chemical data available to make any adjustments for bioavailability (*e.g.*, organic carbon was not measured), which reduces the strength of association and overall weight of the endpoint.

Mussel tissue concentrations are moderate (LOE 3) relative to tissue residue benchmarks, suggesting that the COPCs are bioavailable, are bioaccumulating, and may pose risk to resident shellfish. This measurement has a strong and direct link to predicting effects to the community structure because the benchmark was derived from data that evaluated similar species and environmental exposure conditions.

Direct (toxicity) effects:

Amphipods responded moderately to the sediment contaminants (LOE 2) relative to suitably matched reference sediments. They are the most sensitive organism to the COPCs, and so may not accurately represent impacts to the community as a whole, but do provide a good early-warning indicator that the contamination may be affecting community structure.

Ecological (field) effects:

Benthic community analysis (LOE 4) indicated that the number of taxa at the site was slightly reduced but not significantly lower than at a reference site. In addition, the benthic community richness at the site was reasonably well correlated with chemical concentrations. However, the study was conducted only once and there are other non-chemical stressors that may affect community structure, thereby reducing the strength of this line of evidence.

In conducting the coherence assessment, LOEs should be placed into one of the three LOE categories: measure of exposure, direct measure of toxicity, and ecological effect measure. Using color-coding as shown in Text Box 5 can help visualize what type of information each LOE represents. In the narrative, the results of a given LOE should be

discussed relative to other LOEs within that category, and then discussed relative to LOEs within the other categories to assess whether the ecological effects are related to measures of toxicity and exposure. Correlation analyses and multivariate statistics are tools to evaluate potential relationships between measures of exposure, direct measures of toxicity, and ecological effect measures.

Risk assessors should also communicate the spatial context of the ecological risk to site managers. For example, a few hot spots would be viewed differently than high levels of site-wide contamination. There are valuable techniques the risk assessor can use for clearly representing to the risk manager the spatial aspects of site-associated risks. One technique involves plotting sampling locations on a site map in relation to habitats present at the site (Figure 2). The magnitude of risk for each sample point is indicated by different colors, and the map is visually inspected to ascertain if risk is correlated with habitat type. Another approach is to use a risk zone method (Figure 3). This method is similar to the first, but uses spatial interpolation methods (such as optimal prediction, nearest neighbor interpolation, or kriging) to generate isopleths that represent the spatial extent of risks on a site map. The value of representing the spatial aspects of ecological risk becomes increasingly important as the site size increases.

Figure 2: Habitats in Relation to Risk Zones

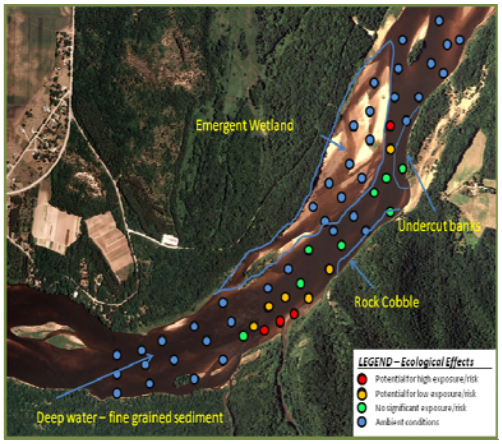
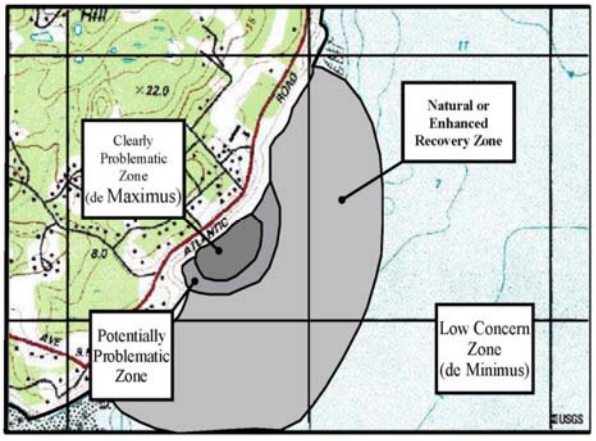


Figure 3: Overview of the Risk Zone Concept



In the narrative, the risk assessor must explain how they will treat the multiple LOEs to achieve a balanced approach to the WOE, particularly if there are different numbers of LOEs in each category (exposure, laboratory, or field-based information) or if one category has significantly higher weighted LOEs than the others. One way to do this is to rely upon the LOE with the highest weight for each category, but the risk assessor should use best professional judgment and an understanding of the ecological system being assessed to determine how to most appropriately combine the LOEs to support a risk conclusion. This coherence analysis is used to explain the ecological relevance of the various endpoints, how they vary with time and space, and how they collectively inform the risk hypotheses. Uncertainties associated with the LOEs, either individually or grouped by category (exposure, laboratory, field) should also be described, along with a quantitative or qualitative discussion of the probability of false positives or negatives. Specifically, the risk narrative should include the following points, at a minimum:

- Clearly stated risk hypotheses;
- Clearly stated assumptions;
- Descriptions of uncertainties associated with each LOE and with the overall risk conclusion;
- Discussion of the probability of Type I (false positive) and Type II (false negative) errors, and the potential consequences of each;
- Discussion of any observed lack of coherence among the LOEs; and
- Final risk conclusion based on the preponderance of the weight of evidence.

Information from relevant peer-reviewed literature should be included in the risk narrative to provide further support for conclusions about risk or causal relationships. Finally, the risk assessor should consider and acknowledge data and associated LOEs that are not available and that could not be considered in the WOE procedure. Uncertainties associated with this lack of information, as well as with LOEs with low final weights, should be clearly described. If a conclusion regarding risk cannot be reached using the available LOEs and/or data, more LOEs and/or data are required.

3.0 REFERENCES

Chapman, P.M., and J. Anderson. 2005. A decision making framework for sediment contamination. *Integr. Environ. Assess. Manage.* 1:163−173.

Hill, A.B. 1965. The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* 58:295-300.

McDonald, B.G., A.M. deBruyn, B.G. Wernick, L. Patterson, N. Pellerin, and P.M. Chapman. 2007. Design and application of a transparent and scalable weight-of-evidence framework: an example from Wabamun Lake, Alberta, Canada. *Integr. Environ. Assess. Manag.* 3(4):476–483.

Menzie, C.A., M.H. Henning, J. Cura, K. Finkelstein, J. Gentile, J. Maughan, D. Mitchell, S. Petron, B. Potocki, S. Svirsky, and P. Tyler. 1996. Special report of the Massachusetts Weight-of-Evidence Workgroup: A weight-of-evidence approach for evaluating ecological risks. *Hum. Ecol. Risk Assess.* 2(2):277–304.

Technical Memorandum

Focused Literature Review of Weight-of-Evidence Approaches

Prepared for

Dr. Paul West
Chair, Science Advisory Board for
Contaminated Sites in British Columbia
Department of Chemistry
University of Victoria
PO BOS 3065 STN C
Victoria, British Columbia V8W 3V6

Prepared by

Exponent
15375 SE 30th Place, Suite 250
Bellevue, WA 98007

November 23, 2009

Doc. no. 0906563.000 01F1 1109 SD23

Literature Review of Weight-of-Evidence Approaches

Introduction

This technical memorandum provides a summary of the focused literature review on weight-of-evidence (WOE) approaches. This review was conducted by Exponent as a preliminary step in the development of guidance for the Science Advisory Board for Contaminated Sites in British Columbia (SABCS) on the use of WOE in ecological risk assessments (ERAs). This memorandum describes the literature search, presents a summary of findings of the focused literature review, and provides an overview of various classification schemes for WOE approaches. Abstracts of the papers and reviewers' notes containing key information from the articles are presented in Appendix A. The citations for the articles discussed in this memorandum are provided in Appendix A, and not provided herein.

Literature Search

The first step was a search of the relevant literature to identify pertinent papers on the use of WOE in ERA. The effort focused on recent literature, and was therefore not intended to be a comprehensive review. The goal was to review the types of WOE methods used in ERAs, the ways in which WOE approaches are classified, and the advantages, disadvantages, and applications of these approaches. A comprehensive review paper by Linkov et al. (2009) served as a starting point for the identification of recent papers on the application of WOE in ERA. In addition, framework papers identified in the recent WOE White Paper (SABCS 2009) were also reviewed.

Focused Literature Review

We reviewed 35 papers that consider various aspects of WOE (Table 1, Appendix A). Several papers provided general reviews of WOE approaches (e.g., Burton et al. 2002; Chapman et al. 2002; Sustainable Fisheries Foundation 2007; Weed 2005; Linkov et al. 2009; Good et al. 1991; Bay et al. 2007), others described WOE frameworks (e.g., Forbes and Calow 2002; Chapman

and Anderson 2005; Smith et al. 2002; Fairbrother et al. 2003), or presented WOE case studies (e.g., McDonald et al. 2007; Critto et al. 2002; Hall et al. 2005; Burkhardt-Holm and Scheurer 2006). Table 1 summarizes the type of papers reviewed (e.g., review paper, case study), the WOE methodology considered (e.g., decision matrix, quantitative), the aim of the study (e.g., identification of causality or impact), the focus of the case study (e.g, sediment or soil), the type of weighting used (e.g., qualitative, quantitative weighting assigned by experts), and provides a brief description of the paper.

Most case studies used best professional judgment along with decision matrices or flowcharts to integrate various lines of evidence (LOEs). Some studies used quantitative approaches, including statistical methods (e.g., ordination, principal components analysis), meta-analysis, scoring, and multi-criteria decision analysis (MCDA). One paper presented a WOE approach that integrated multiple LOEs using response curves that compared contaminated sites to reference sites, with exposure concentrations on the x-axis and divergence from reference conditions on the y-axis (see Lowell et al. 2000). One paper examined the uncertainties associated with LOEs typically considered in sediment quality assessments, but did not present a specific WOE approach (see Bately et al. 2002).

Some studies used WOE to integrate various LOEs and to examine whether a site is impacted (i.e., assess risk); other studies used causal criteria to identify the stressor(s) that contribute to risk. Some WOE approaches clearly describe how LOEs are weighted (e.g., Menzie et al. 1996; Chapman et al. 2002; Chapman and Anderson 2005; McDonald et al. 2007); others use a descriptive rationale (e.g., Forbes and Calow 2004) or causal criteria (e.g., Burkhardt-Holm and Scheurer 2006) to develop weights that are implicit to the development of a decision matrix, while other approaches do not use any weighting of the LOEs (e.g., Smith et al 2002; Reynoldson et al. 2002).

The focus of most studies was on sediment quality (i.e., benthic invertebrate impairment), while a few assessed detrimental effects to fish (e.g., Burkhardt-Holm and Scheurer 2006; McDonald and Chapman 2007; Moraes et al. 2003) and terrestrial wildlife (e.g., Fairbrother 2003). WOE methods for sediment quality assessments appear to be more

prescriptive than other assessments because certain LOE (e.g., sediment chemistry, toxicity tests, and measures of benthic community structure) typically are used in sediment quality assessments.

In summary, the literature review identified WOE papers that ranged from simple approaches (e.g., best professional judgment) to complex quantitative approaches (e.g., MCDA).

Classification of WOE Approaches

Several of the papers reviewed presented classification schemes for the various WOE methodologies (Burton et al. 2002; Linkov et al. 2009; Weed 2005; Sustainable Fisheries Foundation 2007). However, our review indicates a lack of consistency in the WOE classification schemes described below. For instance, the WOE approach described by Menzie et al. (1996) is variously categorized as a consensus ranking system, a scoring system, and a quantitative system (Table 2).

Burton et al. (2002) described eight categories of WOE. The first is qualitative combination, which combines various LOEs in a non-quantitative manner. The second is expert ranking, which uses expert opinions to determine the likelihood of harm. The third is consensus ranking, which involves all stakeholders in the decision-making process regarding risk, based on multiple LOEs (e.g., Menzie et al. 1996). The fourth is semi-quantitative ranking, which includes schemes for pre-screening chemicals, screening threshold effects data (e.g., hazard ranking without exposure considerations), use of scores compiled in an algorithm that weights components, and application of a risk-based quotient method. The fifth is the sediment quality triad, which uses three LOEs (chemical measurements of bulk sediment, laboratory toxicity tests, and field surveys of benthic invertebrate structure). The sixth is tabular decision matrices, in which a decision is made based on the combination of LOE results (e.g., Grapentine 2002). The seventh is broad-scale WOE, in which any method of WOE (e.g., qualitative, ranking ,or quantitative) is used to assess environmental problems that cover large areas (e.g., Lowell et al. 2000). The eighth is quantitative likelihood, which takes numerical data and uses hypothesis testing to reach a conclusion (e.g., Reynoldson et al. 2002; Bailer et al. 2002).

Chapman et al. (2002) divides WOE approaches into two categories: WOE for individual LOEs and WOE that combines multiple LOEs. The first category refers to the integration of multiple endpoints within an overall LOE (e.g., integration of data for various toxicity tests). The second category, in which multiple LOEs are combined, can be further divided into indices, statistical summarization, scoring systems, logic systems, and best professional judgment. The indices approach refers, for example, to the development of ratios of reference values for each LOE in the sediment triad. Statistical summarization includes approaches that use probability values to determine the degree of impact. Scoring systems use a set of attributes to assign the degree of confidence in each LOE used in a WOE. Logic systems are WOE methods that use causal criteria to rule out certain stressors (e.g., Lowell et al. 2000; Forbes and Calow 2002). Best professional judgment uses expert opinion of LOE and available site-specific data to reach a conclusion.

Weed (2005) categorized WOE methods as metaphorical, methodological, or theoretical. Metaphorical WOE methods do not describe or refer to a formal method and lack transparency. Methodological WOE methods examine and intrepret all evidence and include systematic narrative reviews, quality criteria reviews for toxicological studies, epidemiology's causal criteria, meta-analysis, mixed epidemiology-toxicological methods, and quantitative weighting schemes. Theoretical WOE methods include, for example, pattern recognition methods and judiciary methods used in evidentiary gate-keeping. Theoretical WOE methods have no obvious application to ERA.

The Sustainable Fisheries Foundation (2007) provided classification of WOE approaches that could be used to integrate multiple LOEs to assess risk to benthic invertebrates. These included: best professional judgment, tiered approaches, a decision matrix approach (e.g., Bay et al. 2007), a semi-quantitative approach (e.g., Calcasieu Estuary Baseline Ecological Risk Assessment by MacDonald et al. 2002), and a fully quantitative approach (e.g., Menzie et al. 1996).

Linkov et al. (2009) used a combination of categories described by Chapman et al.(2002) and Weed (2005). These include: listing evidence (no integration of LOEs), best professional

judgment, causal criteria (a structured set of criteria for evaluating cause and effect), logic (LOEs are found to either refute, discount, or corroborate a cause or outcome), scoring (weights are assigned to LOE, often using best professional judgment based on various qualities), indexing (weights are assigned to LOEs and integrated into a single value that determines outcome), and quantification (formalized mathematical methods) including MCDA.

Next Steps

Relying on this focused literature review and our experience and expertise, Exponent will develop a conceptual approach to WOE that should be used in detailed ecological risk assessment (DERA). During the problem formulation stage, consideration should be given to various LOEs when developing the site conceptual model and designing studies so that the LOEs have maximum relevance to the assessment endpoints. All stakeholders should agree on which LOEs will be decisional and which will be supportive. During the risk characterization stage, WOE will be used to integrate multiple LOEs for each assessment endpoint to reach a conclusion regarding the potential for risk in a manner that is transpartent and follows relative weightings assigned during Problem Formulation (modified, as appropriate, by quality or quantity changes resulting from actual field conditions). As with other aspects of the DERA, WOE should be standardized but flexible and must be presented in a transparent, understandable manner throughout the DERA.

Exponent will draft a new section on WOE for the existing DERA guidance to introduce WOE and describe when and how it should be used in DERA, building upon the current Section 5.3 and the WOE White Paper written for SABCS (2009). This section will refer the reader to a new technical appendix for details. Exponent will prepare this appendix in the form of a detailed checklist, as a guide to risk assessors on applying WOE in conducting DERAs. Finally, Exponent will provide recommendations for the WOE approach in a meeting with the SABCS in Vancouver, BC.

Tables

Table 1. Summary of weight-of-evidence articles in targeted literature review

Article	Article Type/WOE Type	Study Aim	Assessment Focus	LOE Weighting	Brief Article Description
Alden III et al. (2005)	Case study/Decision Matrix	Impact	Sediment, benthos	Weighting implied in determination of impacts	Integrates results from several previous studies (including Triad studies on sediment contaminants, sediment toxicity, and benthic biological effects; a sediment core study; and a PAH fingerprinting study) to determine impact. Answers “yes” or “no” to a series of key questions, often using statistical and graphical techniques to integrate data.
Bailer et al. (2002)	Case study/Quantitative	Impact	Sediment, benthos	Suggests limiting LOE to only most environmentally relevant	Calculates a site-specific pooled p-value for each test site. Site with smallest pooled p-value considered most affected, largest pooled p-values least affected. Compares plots of pooled p-values for toxicity data versus community structure data to indentify sites with alterations to both.
Batley et al. (2002)	Review/None presented	Impact	Sediment, benthos	-	Discusses uncertainties in WOE .
Bay et al. (2007)	Review/ Best professional judgment	Impact	Sediment, benthos	Various	Compared categorical characterization of 10 experts to a single set of data (e.g., ranging from unimpacted to clearly impacted or inconclusive). Found experts rarely disagreed as to whether site was impacted, more disagreement on magnitude and level of uncertainty. BPJ was a significant source of variation in ranking of sites.
Burkhardt-Holm and Scheurer (2006)	Study/Decision flowchart	Causal	Fish declines	No weighting	Uses causal approach (see Forbes and Calow 2002) and flow chart to assess each possible stressor and categorize each as unlikely, possible, likely, or very likely.
Burton, Chapman et al. (2002)	Review/Various	Impact	Varied	Various	Describes advantages and disadvantages of eight categories of WOE: Qualitative Combination, Expert Ranking, Consensus Ranking, Semi-Quantitative Ranking, Sediment Quality Triad, Broad-Scale WOE, Quantitative Likelihood, and Tabular Matrices.

Table 1. (cont.)

Article	Article Type/WOE Type	Study Aim	Assessment Focus	LOE Weighting	Brief Article Description
Burton, Batley et al. (2002)	Framework/Various	Impact	Various	Various	Describes "certainty elements" that increase reliability of WOE and decision and reduce role of BPJ; selection of critical receptors, defining ecosystem quality, identifying all important stressors and habitat factors that impact receptors and ecosytem quality.
Chapman (2000)	Review/Decision matrix	Impact	Sediment, benthos	Not discussed	Update of SQT method
Chapman (2007)	Framework/Decision matrix	Impact	Sediment, benthos	Weighting considered in developing outcome rules; chemistry assigned least weight, community structure most weight	Uses symbols to populate decision matrix, provides outcomes ranging from negligible risk to high risk. Considers effluent investigations.
Chapman (2007)	Editorial/Various	-	No detailed case studies or examples	Various	Compares and contrasts traditional ecological knowledge of indigenous people with scientific weight of evidence approach. Advocates consideration of traditional ecological knowledge.
Chapman and Anderson (2005)	Framework/Decision matrix	Impact	Sediment, benthos	Weighting considered in developing outcome rules; chemistry assigned least weight, community structure most weight	Provides flowchart and decision matrix. Uses symbols to populate matrix, provides outcomes ranging from no further action to management action required.
Chapman et al. (2002)	Review and Framework/Decision matrix	Impact	Sediment, benthos	Weighting considered in developing outcome rules	Provides flowchart and decision matrix. Uses symbols to populate matrix, provides outcomes ranging from negligible effects to significant effects.
Critto et al. (2007)	Case study/Quantitative	Selection of LOE	Soil	Assigned by system experts	Uses MCDA to identify LOE and develop a tiered ERA framework.
Fairbrother (2003)	Framework/Decision flowchart	Impact	Wildlife	Tiered approach with lower tiers have less weight and higher tiers have more weight	Develops tiered approach and flowchart; describes uncertainties in lower tiers
Forbes and Calow (2002)	Framework/Decision matrix	Causal	Various examples	Implied in development of matrix of causation outcomes	Presents seven questions, and flow diagram and matrix for assigning causation as unlikely (or possibly if data are poor), likely, or very likely.

Table 1. (cont.)

Article	Article Type/WOE Type	Study Aim	Assessment Focus	LOE Weighting	Brief Article Description
Forbes and Calow (2004)	Case study/Decision matrix	Causal	Sediment, benthos	Implied in development of matrix of causation outcomes	Presents seven questions, and flow diagram and matrix for assigning causation to each stressor as unlikely (or possibly if data are poor), likely, or very likely
Good (1991)	Review/None presented	-	-	-	General discussion of WOE theory.
Grapentine, Anderson et al. (2002)	Framework/Decision matrix	Impact	Sediment, benthos	Implied in development of matrix	Uses +/- for each LOE, provides flow diagram and matrix of outcomes, ranging from no risk to adverse effects that require management.
Grapentine, Marvin et al. (2002)	Case study/Decision matrix	Impact	Sediment, benthos	Based on scope, frequency, and amplitude	Addresses integration of data within sediment chemistry LOE
Hall et al. (2005)	Case study/Decision matrix	Impact and Causal	Sediment, benthos	-	Presents various LOE. WOE presented in Alden et al. (2005).
Krimsky (2005)	Review/Various	-	No detailed case studies or examples	Various	Notes recent court case that could be grounds for judicial dismissal of WOE methods.
Linkov et al. (2009)	Review/Various	Various	No detailed case studies or examples	Various	Defines categories of WOE: listing evidence (no integration of LOE); BPJ; causal criteria (presents a structure for evaluating cause and effect); logic (identifies LOE that either refute, discount, or corroborate a cause or outcome), scoring (assigns weights to LOE, often using BPJ based on various qualities, such as strength of assoc, etc., weights for individual LOE are combined into a numerical score); indexing (assigns weights to LOE and integrates LOE into a single value that determines outcome); quantification (uses formalized mathematical methods). Note that neither indexing nor scoring use formal decision analysis techniques, so transparency and reproducibility are limited.
Lowell et al. (2000)	Case study/Graphical display of response curves	Causal	Various aquatic LOE	Each LOE evaluated in terms of causation criteria; weighting based on magnitude of exceedance of critical threshold; no detailed description of weighting provided	Presents integrated response curves communities at contaminated sites relative to reference communities, with concentration on x-axis and divergence from reference on y-axis.

Table 1. (cont.)

Article	Article Type/WOE Type	Study Aim	Assessment Focus	LOE Weighting	Brief Article Description
McDonald and Chapman (2007)	Case study/Decision matrix	Impact	Fish impacts, selenium	Implied in development of matrix; chemistry assigned least weight	Tiered approach. LOE range from comparison of tissue concentration to toxicity reference values to reproductive toxicity tests on field-collected eggs and fish population assessment.
McDonald et al. (2007)	Case study/Decision matrix	Impact	Sediment, benthos	Used both a priori and a posteriori weighting factors	Used 38 LOE. Transparency and reduction in potential influence bias were main goals.
McPherson et al. (2008)	Case study/Decision matrix	Impact	Sediment, benthos	Weighting considered in developing outcome rules	Examined influence of water depth as separate LOE for benthic community structure.
Menzie et al. (1996)	Framework/Decision matrix	Impact	Various	Weighting done qualitatively or quantitatively based on 11 attributes	Attributes are strength of association, site specificity, stressor specificity, data quality, availability of objective criteria/indices, sensitivity, spatial representativeness, temporal representativeness, quantitativeness, stressor response correlation, and use of a standard method. Uses matrix to illustrate magnitude and weight of LOE. WOE can be quantitative or qualitative.
Moraes et al. (2003)	Case study/Decision matrix	Impact and Causal	Fish, metals	Uses causal criteria: strength, gradient, consistency, plausibility, and specificity	LOE range from tissue concentrations to community structure. Relies on establishing causal relationship between LOE and effects. Develop matrix to define whether LOE are/are not consistent with effect or are ambiguous.
Reynoldson, Smith et al. (2002)	Case study/Quantitative	Impact	Sediment, benthos	No weighting	Uses three methods to compare large data set of reference sites to contaminated sites: multivariate clustering, meta-analysis (i.e., calculates p-values to designate level of impact), uses probabilites/odds ratios. Results vary among methods, no recommendation of method provided.
Reynoldson, Thompson et al. (2002)	Case study/Quantitative	Impact	Sediment, benthos	One scoring method uses weighting, others do not	Uses three methods to compare large data set of reference sites (n=222 sites) to contaminated sites: two scoring methods and multivariate statistics (ordination). Recommends ordination as least subjective and most subjective.
Semenzin et al. (2007)	Case study/Quantitative	Selection of LOE	Soil	Assigned by system experts	Uses MCDA to identify bioavailability tools in a tiered ERA.

Table 1. (cont.)

Article	Article Type/WOE Type	Study Aim	Assessment Focus	LOE Weighting	Brief Article Description
Semenzin et al. (2008)	Case Study/Quantitative	Impact	Soil	Assigned by system experts	Uses MCDA to integrate various LOE and calculate integrated effect indexes.
Smith et al. (2002)	Framework/Quantitative	Impact	Sediment, benthos	No weighting	Uses probabilities/odds ratios. Decision made on the basis of the likelihood.
Sustainable Fisheries Foundation (2007) for BC MOE	Review/Various	-	Sediment, benthos	Various	Classification of candidate WOE approaches included: BPJ, tiered approaches, decision matrix approach, semi-quantitative approach, and fully quantitative approach. Presents list of desirable characteristics: supportive of management decisions, scientifically defensible, consistent with narrative intent, consistent with uncertainty assessment, reproducible, transparent, and reliable. Approaches must identify threshold of acceptable and unacceptable (as defined in DERA) risk.
Weed (2005)	Review/Various	Various	Various	Various	Describes and provides examples: metaphorical, methodological, and theoretical. Summarizes issues of concern related to variability and inconsistency.

Note: DERA - detailed ecological risk assessment
ERA - ecological risk assessment
LOE - line of evidence
MCDA - multi-criteria decision matrix
SQT - sediment quality triad
WOE - weight of evidence

Table 2. Classification schemes for weight-of-evidence approaches

- **Categories from Burton et al. (2002)[a]**
  - Qualitative Combination
  - Expert Ranking
  - Consensus Ranking (e.g., Menzie et al. 1996)
  - Semi-Quantitative Ranking
  - Sediment Quality Triad
  - Tabular Decision Matrices
  - Broad-Scale WOE
  - Quantitative Likelihood
- **Categories from Chapman et al. (2002)[b]**
  - Individual Lines of Evidence
  - Combined Lines of Evidence
    - Indices
    - Statistical summarization
    - Logic systems
    - Scoring systems (e.g., Menzie et al. 1996)
    - Best Professional Judgment
- **Categories from Weed et al. (2005)[c]**
  - Metaphorical (no method described)
  - Methodological
    - WOE method versus a “strength of evidence” approach
    - WOE method using “all” rather than a selected subset (e.g., standard test assay) of the evidence
    - WOE method pointing to other “established” or familiar interpretative methodologies
      - Systematic narrative review
      - Quality criteria for toxicological studies
      - Epidemiology’s causal criteria
      - Meta-analysis
      - Mixed epidemiology-toxicology methods
    - WOE method employing a quantitative weighting scheme (e.g., Menzie et al. 1996)
  - Theoretical method (e.g., pattern recognition in cognitive science)

Table 2. (cont.)

Categories from Sustainable Fisheries Foundation, *Workshop Summary Report,* December 2007[d]
Best professional judgment approach
Tiered approaches
Decision matrix approach (consistent with the approach used in California; Bay et al. 2007)
Semi-quantitative approach (consistent with the approach used in the Calcasieu Estuary; MacDonald et al. 2002)
Fully quantitative approach (e.g., Menzie et al. 1996)
Categories from Linkov et al. (2009)[e]
Listing Evidence
Best Professional Judgment
Causal Criteria
Logic
Scoring
Indexing
Quantification

Note: WOE - weight of evidence

[a] Burton, A.G.J., P.M. Chapman, and E.P. Smith. 2002. Weight-of-evidence approaches for assessing ecosystem impairment. Hum. Ecol. Risk Assess. 8(7):1657−1673.

[b] Chapman, P.M., B.G. McDonald, and G.S. Lawrence. 2002. Weight-of-evidence issues and frameworks for sediment quality (and other) assessments. Hum. Ecol. Risk Assess. 8(7):1489−1515.

[c] Weed, D.L. 2005. Weight of evidence: A review of concept and methods. Risk Anal. 25(6):1545−1557.

[d] Sustainable Fisheries Foundation. 2007. Workshop to support the development of guidance on the assessment of contaminated sediments in British Columbia: Workshop summary report. Prepared for B.C. Ministry of Environment, Land Remediation Section. December, 2007.

[e] Linkov, I., D. Loney, S. Cormier, F.K. Satterstrom, and T. Bridges. 2009. Weight-of-evidence evaluation in environmental assessment: Review of qualitative and quantitative approaches. Sci. Total Environ. doi:10.1016/j. scitotenv.2009.05.004.

Appendix A

Abstracts and Reviewers' Notes

Exponent®

A Review of Selected Literature on Weight-Of-Evidence

From "Published Frameworks" Noted in June 17, 2009, SAB White Paper

Bailer, A. J.; Hughes, M. R.; See, K.; Noble, R.; Schaefer, R., A Pooled Response Strategy for Combining Multiple Lines of Evidence to Quantitatively Estimate Impact. *Human Ecol. Risk Assess.* 2002, *8*, (7), 1597-1611.

Abstract (reproduced from article)

The impacts of sediment contaminants can be evaluated by different lines of evidence, including toxicity tests and ecological community studies. Responses from 10 different toxicity assays/tests were combined to arrive at a "site score." We employed a relatively simple summary measure, pooled *P*-values where we quantify a potential decrement in response in a contaminated site relative to nominally clean reference sites. The response-specific *P*-values were defined relative to a "null" distribution of responses in reference sites, and were then pooled using standard meta-analytic methods. Ecological community data were also evaluated using an analogous strategy. A distribution of distances of the reference sites from the centroid of the reference sites was obtained. The distance from each of the test sites from the centroid of the reference sites was then calculated, and the proportion of reference distances that exceed the test site difference was used to define an empirical *P*-value for that test site. A plot of the toxicity P-value versus the community *P*-value was used to identify sites based on both alteration in community structure and toxicity, that is, by weight-of-evidence. This approach provides a useful strategy for examining multiple lines of evidence that should be accessible to the broader scientific community. The use of a large collection of reference sites to empirically define *P*-values is appealing in that parametric distribution assumptions are avoided, although this does come at the cost of assuming the reference sites provide an appropriate comparison group for test sites.

Notes

Used the following methodology:

1. Calculate a site-specific pooled *P*-value for each test site for both toxicity data and benthic community structure data
2. Order sites based on these pooled values
3. Sites with the smallest pooled *P*-value are considered most affected. Sites with largest pooled *P*-value are considered least affected
4. Simple graphical assessments are examined (plot of benthic community *P*-value versus sediment toxicity test *P*-value).

Categorizes sites as: Most affected (P-value $\leq$ 0.05); Moderately affected (0.05 < P-value < 0.10); Minimally affected (P-value $\geq$ 0.10). Critical issue is the formation of an appropriate data set for reference site. Case study presented used 146 reference sits and 66 test sites. Overall method is relatively simple, accessible, and fairly easy to implement.

Batley, G. E.; Burton, G. A.; Chapman, P. M.; Forbes, V. E., Uncertainties in Sediment Quality Weight-of-Evidence (WOE) Assessments. *Human Ecol. Risk Assess.* 2002, *8*, (7), 1517-1547.

Abstract (reproduced from article)

Uncertainties in sediment quality assessments are discussed in five categories: (1) sediment sampling, transport and storage; (2) sediment chemistry; (3) ecotoxicology; (4) benthic community structure; and (5) data uncertainties and QA/QC. Three major exposure routes are considered: whole sediments, and waters in sediment pores and at the sediment-water interface. If these uncertainties are not recognized and addressed in the assessment process, then erroneous conclusions may result. Recommendations are provided for addressing the identified uncertainties in each of the key areas. The purpose of this paper is to improve the reporting of sediment quality assessments.

Notes

This paper presents uncertainties associated with each LOE for the sediment triad approach. It is useful for assigning weights to LOEs for sediment assessments, but doesn't provide a WOE approach.

Bay, S.; Berry, W.; Chapman, P. M.; Fairey, R.; Gries, T.; Long, E.; MacDonald, D.; Weisberg, S. B., Evaluating Consistency of Best Professional Judgment in the Application of a Multiple Lines of Evidence Sediment Quality Triad. *Integrated Environ. Assess. Manag.* 2007, *3*, (4), 491–497.

Abstract (reproduced from article)

The bioavailability of sediment-associated contaminants is poorly understood. Often, a triad of chemical concentration measurements, laboratory sediment toxicity tests, and benthic infaunal community condition is used to assess whether contaminants are present at levels of ecological concern. Integration of these 3 lines of evidence is typically based on best professional judgment by experts; however, the level of consistency among expert approach and interpretation has not been determined. In this study, we compared the assessments of 6 experts who were independently provided data from 25 California embayment sites and asked to rank the relative condition of each site from best to worst. The experts were also asked to place each site into 1 of

6 predetermined categories of absolute condition. We provided no guidance regarding assessment approach or interpretation of supplied data. The relative ranking of the sites was highly correlated among the experts, with an average correlation coefficient of 0.92. Although the experts' relative rankings were highly correlated, the categorical assessments were much less consistent, with only 1 site out of 25 assigned to the same absolute condition category by all 6 experts. Most of the observed categorical differences were small in magnitude and involved the weighting of different lines of evidence in individual assessment approaches, rather than interpretation of signals within a line of evidence. We attribute categorical differences to the experts' use of individual best professional judgment and consider these differences to be indicative of potential uncertainty in the evaluation of sediment quality. The results of our study suggest that specifying key aspects of the assessment approach a priori and aligning the approach to the study objectives can reduce this uncertainty.

Notes

There were considerable differences in how the experts rated the sites. Each expert used a different specific approach based on respective philosophy and experience. Some experts used a numeric approach that integrated scores or ranks based on levels of response within a line of evidence (LOE), whereas others based their clasifications on more subjective comparisons of concordance and relative magnitude among the LOEs. All experts agreed that it was critical to demonstrate a linkage between chemical exposure and biological effects.

Several experts felt that complete chemistry data sets were lacking and thought including data for additional chemicals would have given them more confidence in their evaluation. The availability of only a single toxicity test and lack of sublethal endpoints concerned the experts. Field benthic community data were important to the experts in their evaluations, but many cited uncertainties in interpreting the data and issues related to distinguishing contaminant effects from effects related to habitat or physical factors.

The experts thought that the predetermined categories defined in this study were too ambiguous, thus leading to differences in intrepretation.

The experts rarely disagreed about whether a site was impacted or unimpacted, but disagreed on the magnitude and certainty of the impact. The use of best professional judgment in the WOE approach was found to be a significant source of variation in the evaluation and ranking of the sites. This in turn could lead to uncertainty in the risk conclusions and affect how the site is managed. The study recommended several steps to reduce such uncertainty in intrepretation of sediment quality triad data. First, the relative weight of each LOE, how multiple LOEs will be combined, and the criteria for determining the assessment conclusion should be determined during the study design. Second, guidance on specfiic methods for measuring sediment chemistry, toxicity, and benthic community condition could improve comparability among site assessments. Third, uncertainty in sediment quality assesments can be reduced by training the individuals interpreting the data.

Burton, Jr., G.A.; Batley, G. E.; Chapman, P. M.; Forbes, V. E.; Smith, E. P.; Reynoldson, T.; Schlekat, C. E.; Besten, P. J. d.; Bailer, A. J.; Green, A. S.; Dwyer, R. L., A Weight-of-Evidence Framework for Assessing Sediment (Or Other) Contamination: Improving Certainty in the Decision-Making Process. *Human Ecol. Risk Assess.* 2002, *8*(7): 1675–1696.

Abstract (reproduced from article)

A basic framework is presented for the ecological weight-of-evidence (WOE) process for sediment assessment that clearly defines its essential elements and will improve the certainty of conclusions about whether or not impairment exists due to sediment contamination, and, if so, which stressors and biological species (or ecological responses) are of greatest concern. The essential "Certainty Elements" are addressed in a transparent best professional judgment (BPJ) process with multiple lines-of-evidence (LOE) ultimately quantitatively integrated (but not necessarily combined into a single value). The WOE Certainty Elements include: (1) Development of a conceptual model (showing linkages of critical receptors and ecosystem quality characteristics); (2) Explanation of linkages between measurement endpoint responses (direct and indirect with associated spatial/temporal dynamics) and conceptual model components; (3) Identification of possible natural and anthropogenic stressors with associated exposure dynamics; (4) Evaluation of appropriate and quantitatively based reference (background) comparison methods; (5) Consideration of advantages and limitations of quantification methods used to integrate LOE; (6) Consideration of advantages and limitations of each LOE used; (7) Evaluation of causality criteria used for each LOE during output verification and how they were implemented; and (8) Combining the LOE into a WOE matrix for interpretation, showing causality linkages in the conceptual model. The framework identifies several statistical approaches for integrating within LOE, the suitability of which depends on physical characteristics of the system and the scale/nature of impairment. The quantification approaches include: (1) Gradient (regression methods); (2) Paired reference/test (before/after control impact and ANOVA methods); (3) Multiple reference (ANOVA and multivariate methods); and 4) Gradient with reference (regression, ANOVA and multivariate methods). This WOE framework can be used for any environmental assessment and is most effective when incorporated into the initial and final study design stages (*e.g.*, the Problem Formulation and Risk Characterization stages of a risk assessment) with reassessment throughout the project and decision-making process, rather than in a retrospective data analysis approach where key certainty elements cannot be adequately addressed.

Notes

This paper presents a framework for ecological WOE approaches to assessing sediment contamination, but it could be applied to water and soil quality assessments. "Certainty Elements" are identified to help increase reliability of WOEs and decision-making and reducing the role of BPJ. The Certainty Elements include selection of **critical receptors (species)**, definition of **ecosystem quality** (as it will have an impact on how reference sites are selected), and identification of **all stressors and habitat factors that could affect the critical species or ecosystem quality**. Relationships between stressor and receptors should be understood spatially and temporally. The next step of the framework is to develop a conceptual model that links the

critical receptors, ecosytem quality characteristics and the stressors (natural and anthropogenic). Then measurement endpoints are selected based on the conceptual model and the strengths and limitations of each measurement endpoint (Table 2 provides a nice summary of advantages and limitations of several LOEs used in sediment assessments). In Reference comparisons are made and reference site selection is critical. Then a study design and QA/QC plan is developed to identify data quality objectives. LOEs are analyzed in terms of QA/QC; stressor magnitude, frequency, duration and interactions, and exposure-biological response relationships. Causality criteria used to link stressors and biological responses should be clearly stated and should be based on spatial correlation, temporal correlation, strength of link, consistency of association, experimental confirmation (lab or field), plausibility, and specificity. The final step is to integrate the LOEs into a WOE matrix using Koch's postulates (Suter 1993): adverse effect must be regularly associated with exposure to stressor, the stressor must be found in the receptor, the adverse effect must be manifested in unimpaired species following exposure under controlled exposure conditions, and there must be an indicator of exposure. The WOE matrix should show causality linkage in the conceptual model.

This paper doesn't describe a specific WOE approach for analyzing and weighting the LOEs. It states that results of the expert judgments can be summarized in a decision matrix, converting to ranks using numbers or +/-. Possible approaches include ranking by uncertainty, ecological relevance, or societal importance.

How will Type I (false positive) and Type II (false negative) errors be addressed in the decision process? Typically, the Type I error rate is kept at a low level in statistical analysis and sample size is used to control the Type II error rate.

This paper states that combining the information in multiple LOEs into a single number that describes degree of impairment results in excessive reduction of information and over-simplifies the evidence. Also, expert judgment must be carefully incorporated and well documented to ensure transparency. The greatest subjectivity occurs during the determination of causality, when the LOE results are analyzed in the final WOE matrix.

Burton, Jr., G.A.; Chapman, P. M.; Smith, E. P., Weight-of-Evidence Approaches for Assessing Ecosystem Impairment. *Human Ecol. Risk Assess.* 2002, *8*, (7), 1657-1673.

Abstract (reproduced from article)

It is challenging determining whether an ecosystem is impaired. The complexity of direct and indirect interactions between physical, biological and chemical components with their varying temporal and spatial scales generally renders use of multiple assessment approaches mandatory, with a consequent need to integrate different lines-of-evidence. Integration generally involves some form of weight-of evidence (WOE). WOE approaches reported in the literature vary broadly from subjective and qualitative to quantitative. No standard approach exists and no accepted guidelines exist describing how a WOE process should be conducted. This review summarizes the advantages, limitations, and uncertainties of different WOE approaches, critical

issues involved in selecting and executing different lines-of evidence, and the process for subsequent characterization of the likelihood of impairment.

Notes

WOE defined as "the process of combining information from multiple LOE to reach a conclusion about and environmental system or stressor" and also a process that "incorporates judgments about the quality, extent and congruence of the information in each LOE."

- "…the environmental sciences require a formalization of the WOE process, in other words a structuring of the process to reduce the liklihood of bias and increase the liklihood that accurate predictions, correct conclusions, and/or proper decisions will be made."
- "Ideally, this process would include: formulation of the technical question(s) to be answered and associated facts; selection of experts; training of experts regarding process and judgmental biases; decomposition of the technical question and clear definition of the variables or other inputs; elicitation of probability distributions from individual experts; aggregation and discussion of indiviudal differences among experts; processing the information, documentation (from substantive and normative experts) and communication (Peters et al. 1998). However, the above ideal process is rarely the reality due to limitations in both time and resources."

Interesting summary of LOE: For sudies of ecosystem impairment, the LOEs used in the WOE proces have included: 1) Comparison of site chemistry to individual chemical values (e.g, criteria, background reference); 2) Comparison of indigenous populations between site and reference; 3) comparison of lab toxicity tests between site, reference, and controls; 4) Comparison of chemical tissue residues in indigenous biota to reference sites or literature values; 5) Evaluation of model predictions of fate and/or effects (e.g., contaminant transport, bioaccumulation) at the site compared to a gradient, reference sites, or literature values. In most studies, only 2 or 3 LOEs. Most common are chemistry, indigenous biota, and lab toxicity.

Selecting and effectively impementing the propoer combination of assessment approaches are essential to impairment assessments. Proper selection and execution of multiple LOEs, and the subsequent characterization and combination of the various LOEs into a WOE-based decision are inextricably linked.

Various WOE methodologies are examined in terms of strengths and weaknesses of each WOE approach. Categories include: robustness (consistency in interpretation and decision-making irrespective of when and where conducted), methodology (ease of use), sensitivity (how well the approach can discriminate among levels of effects, from small to extreme effects), appropriateness and application (whether the approach is useful for a wide range of conditions or environments), and transparency (how understandable the approach is).

Examines and evaluates 8 approaches:

Qualitative Combination or "Lumping" various LOEs in a nonqualitative manner. Example: toxicity in 2 of 3 tests leads to the conclusion that he sample is toxic. Number and magnitude of exceedances of chem criteria indicates that it is more impaired than a site with fewer, smaller exceedances.

Ranking Approaches:

Expert Ranking. Mostly based on expert opinion regarding priority characteristics. Typically relies heavily on chemistry.

Consensus Ranking. Stakeholders define the approach as part of problem formulation, depends on strength of association, data quality, study design, and execution attributes. Menzie et al 1996. Note that this WOE approach gives greater weight to endpoint concurrence, which this author considers problematic. Lack of concurrence among various LOEs does not necessarily mean an LOE is innaccurate. Advantage is that you can achieve consensus among stakeholders on study design and interpretation, *a priori.* Disadvantgage is that BPJ-weighting varies in quality and accuracy depending on the expertise of the stakeholders. Training of stakeholders can be useful. This approach is similar to expert ranking, but has a higher level of transparency.

Semi-qualitative Ranking. Different LOE data can be normalized (e.g., to percentiles), ranked, and evaluated in tandem. Some use bivariate and stepwise linear regression to select the best ten LOEs for a model. Often need to adjust various metrics for certain non-stressor variables (e.g., fish metrics are adjusted for stream size, community structure adjusted for salinity).

Sediment Quality Triad. Data from each LOE were initially normalized to reference (as a percentage) and presented graphically as a triangle. However, because of information loss in the presentation it has been replaced by more quantitative approaches.

Tabular Decision Matrices. Advantage is ability to rapidly disseminate the final WOE finding. Example is Grapentine et al. 2002. Matrices typically provide info on individual LOEs in a binary classification: toxic or not, contaminated or not, altered or not. Greater levels of discrimination are possible but become more subjective. Backup tables on individual LOE can be used.

Broad-Scale WOE. Incorporates varous WOE approaches. Uses knowledge gained from a variety of sites with similar stressors. Critical to the process and to establishing causality is a high quality knowledge base (i.e., expertise + data). Causality established using criteria from epidemiology (e.g., Fox 1991; Hill 1965, etc.) in which some criteria given greater weight than others. Provides 2 sets of causality criteria.

- Lowell et al. 2000. Spatial and/or temporal correlations of stressor and effect; plausible mechamisms for stressor and effect link; experimental

verification of link; strength and specificity of link; biomarker evidence; consistency of link over large geographic area; coherence of link in other regions. Effect limits (e.g., critical effect size) are determined to establish the level at which an LOE is judged ecologically significant.

- EPA 2000 and Suter 2002. Nonquantitative stressor ID evaluation process. Set of 4 basic criteria, which are evaluated using 10 causal evaluations: 1) co-occurrence, 2) temporality, 3) biological gradient, 4) complete exposure pathway, 5) consistency of association, 6) experimental confirmation, 7) plausibility, 8) specificity, 9) analogy, 10) predictive performance. These BPJ evaluations are summarized in a table, converted to ranks (e.g, 1 to 4, or +/-).

Quantitative Likelihood. Based on hypothesis testing (e.g., impaired versus not impaired) and WOE through likelihood. Approach often leads to misinterpretations: users interpret probabilities as the probability that the hypothesis is true, rather than the probability of observing the evidence given that the hypothesis is true. Other problems relate to sample size: 2 studies with similar *P*-value may have different interpretations if one has large n, and one small n. A related approach calculates how likely the data are under the different hypotheses, uses a likelihood ratio test, e.g., if the data are much more likely under the null than the alternative hypothesis, then the ratio of alternative to null will be small. Difficulties arise when trying to specify hypotheses and error rates, because of the need to define states of impact/no impact. Advantage is that it provides a means for directly combining information to estimate an impairment probability, allows for the combination of *P*-values from multiple LOEs to a single LOE (e.g., PCA for multiple chemicals, multivariate methods, also biotic indices, but they may result in information compression and loss of information). Also includes Species Sensitivity Distributions.

Chapman, P. M.; McDonald, B. G.; Lawrence, G. S., Weight-of-Evidence Issues and Frameworks for Sediment Quality (And Other) Assessments. *Human Ecol. Risk Assess.* 2002, *8*, (7), 1489-1515. (Special edition of HERA on WOE assessment.)

Abstract (reproduced from article)

Weight of evidence (WOE) frameworks for integrating and interpreting multiple lines of evidence are discussed, focusing on sediment quality assessments, and introducing a series of ten papers on WOE. Approaches to WOE include individual lines of evidence (LOE) as well as combined LOE (indices, statistical summarization, logic systems, scoring systems, and best professional judgment [BPJ]). The application of WOE, based on multiple LOE, is discussed relative to the published literature. Fully implementing WOE requires consideration of six main LOE in sediment (or other assessments); these LOE generally correspond to other causality considerations including Koch's Postulates. However, the issue of sediment stability is an additional consideration, and the use of tabular decision matrices is recommended in a logic system to address LOE described by others as "analogy", "plausibility", or "logical and

scientific sense." Three examples of logic system WOE determinations based on the Sediment Quality Triad and using tabular decision matrices are provided. Key lessons from these examples include the: generally limited utility of sediment quality value (SQV)-based LOE; need for BPJ; importance of ecological relevance; importance of assessing background conditions; and, need for appropriately customizing study designs to suit site specific circumstances (rather than application of "boiler-plate" assessments). Overall, more quantitative approaches are needed that better define certainty elements of WOE in an open framework process, *i.e.*, statistical summarization culminating in a logic system incorporating BPJ.

Notes

Paper serves as introduction to a series of WOE papers in the journal, *HERA*. WOE defined as "a determination related to possible ecological impacts based on multiple lines of evidence." Summarizes various WOE approaches.

- **Indices**. Sediment Quality Triad (SQT) was originally based on indices, the ratio to reference (RTR) values for chemistry, toxicity, and benthic community structure. However, use of indices can lead to information compression. For biologial data, which are direct measures of adverse effects, there is no need to further reduce the data. However, efforts to develop indices continue and examples of various indices are provided.

- **Statistical Summarization**. One approach uses probability elipses of clustered data to determine the difference from reference conditions. Another approach estimates the probability of environmental harm using Bayes Theorem, based on the odds ratio (the likelihood of two different scenarios – impacted, or not impacted). Another approach uses meta-analysis, where *P*-values and effect sizes are pooled. Common issues related to these appraoches incude: definition of appropriate reference conditions, and defining "impact."

- **Scoring Systems**. Presents example of Menzie et al. (1996). Measurement endpoints are weighted by stakeholders using 10 separate judging attributes. Results are presented in a tabluar decision matrix. Can also be conducted qualitatively.

- **Logic Systems**. Likely originated with Koch's (1984) postulates. Also applied to SQT. Provides examples of various use of logic in WOE assessments to determine impact or cause.

- **Best Professional Judgment**. Defined as the use of expert opinion and judgment based on available data and site- and situation-specific conditions to determine environmental status or environmental risk. The Precautionary Principle can be considered as BPJ. Also, statistics can involve use of BPJ, as in selection of *P*-value.

Discusses need to acknowledge and address uncertainty in WOE applications. For example, uncertainty in WOE sediment assessments fall into five categories: 1) sediment sampling, transport and storage, 2) sediment chemistry, 3) ecotoxicology, 4) benthic community structure, and 5) data uncertainties and QA/QC.

Provides a table that compares six LOEs for sediment assessments to seven LOEs proposed for retrospective assessments (Forbes and Calow 2002) and to three other sets of causality considerations that reflect Koch's postulates. Although LOEs are similar, various authors use different approaches to combine and summarize them.

Provides three examples of WOE assessments that use tabular decision matrices as the basis for sediment management decision-making. Concludes by noting the importance of identifying a transparent and quantitative process for weighting the LOE. Advocates for use of a statistical summary that is incorporated into a logic system.

Forbes, V. E.; Calow, P., Applying Weight-of-Evidence in Retrospective Ecological Risk Assessment When Quantitative Data Are Limited. *Human Ecol. Risk Assess.* 2002, *8*, (7), 1625-1639.

Abstract (reproduced from article)

Retrospective ecological risk assessment attempts to identify likely causal agents to explain adverse effects that have occurred in ecological targets. It can never be decisive since it is *post hoc* and usually based on limited evidence that is rarely very quantitative. It can, nevertheless, be made more transparent, systematic, and logical, and less subjective. Based on human health epidemiological criteria we develop an approach that moves from systematic consideration of seven basic questions to assigning a likelihood of involvement of putative agents. The questions are: 1. Is there evidence that the target is or has been exposed to the agent? 2. Is there evidence for correlation between adverse effects in the target and exposure to the agent either in time or in space? 3. Do the measured or predicted environmental concentrations exceed quality criteria for water, sediment or body burden? 4. Have the results from controlled experiments in the field or laboratory led to the same effect? 5. Has removal of the agent led to amelioration of effects in the target? 6. Is there an effect in the target known to be specifically caused by exposure to the agent? 7. Does the proposed causal relationship make sense logically and scientifically? We identify 15 common scenarios of answers to the questions and illustrate the approach by reference to three real-world case studies (decline in benthos in a tropical marine bay, decline in fisheries in a temperate sea, decline in marine mollusc populations). The primary challenge in retrospective risk assessment is to make best use of the available evidence to develop rational management strategies and/or guide additional analyses to gain further evidence about likely agents as causes of observed harm.

Notes

Advocates that any approach should be systematic, transparent, and logical. Presents series of questions, as noted in Abstract. Possible answers include *no, yes*, and *no data*. Provides a flowchart that translates the various combinations of answers to the questions into conclusions about relative likelihood that the identified agents cause the observed effects. Levels of likelihood expressed as conclustions include: *don't know, unlikely, possibly, likely,* and *very likely*. Provides three example case studies.

Grapentine, L.; Anderson, J.; Boyd, D.; Burton, G. A.; DeBarros, C.; Johnson, G.; Marvin, C.; Milani, D.; Painter, S.; Pascoe, T.; Reynoldson, T.; Richman, L.; Solomon, K.; Chapman, P. M., A Decision Making Framework for Sediment Assessment Developed for the Great Lakes. *Human Ecol. Risk Assess.* 2002, *8*, (7), 1641-1655.

Abstract (reproduced from article)

A rule-based, weight-of-evidence approach for assessing contaminated sediment on a site-by-site basis in the Laurentian Great Lakes is described. Information from four lines of evidence — surficial sediment chemistry, laboratory toxicity, invertebrate community structure and invertebrate tissue biomagnification — is integrated within each line to produce a pass ('–') or fail ('+') conclusion, then combined across lines resulting in one of 16 outcome scenarios. For each scenario, the current status of the site, interpretation, and management recommendations are given. Management recommendation(s) can range from no action to risk management required (9 of the 16 scenarios). Within each line of evidence, the strength of each response can also be ranked (*e.g.*, score of 1 to 4), providing managers with more information to aid decision options. Other issues that influence scientific management recommendations include site stability, subsurface contamination and spatial extent of effects. The decision framework is intended to be transparent, comprehensive (incorporating exposure, effect, weight-of-evidence, and risk), and minimally uncertain.

Notes

This study reports a rule-based WOE approach for assessing contaminated sediments. Four lines of evidence are used (sediment chemistry, laboratory toxicity tests, invertebrate community structure and invertebrate tissue biomagnification) and with each line yielding a pass (-) or fail (+). The pass/fail conclusions for the LOEs are combined and result in one of 16 possible scenarios, some of which indicate risk, others no risk, and many indicating a risk management decision is required and/or more evaluation is needed.

The first step is to synthesize the data and determine pass/fail – Reaching pass/fail should be based on clearly stated stastical criteria. Statistical ordination is favored for characterizing sites in terms of 2 or 3 variables and assessing differences between test and reference sites. A significant effect is indicated when conditions in a test site fall outside a 95% confidence limit for reference sites. This could lead to Type I and or Type II errors and thus the limit can be

adjusted. For sediment chemistry, one can use PCA to compare test sites to reference sites, develop hazard quotients, use Persaud et al. (1993) screening levels, or develop sediment quality index based on the Canadian Water Quality Index. For benthic community structure, the test site and reference site samples are plotted on the same ordination space and the community is ranked from unaltered to severly altered based on distance from test site to reference sites and then given pass or fail designations. For sediment toxicity, statistical differences in endpoints between test sites and reference sites are used to assign pass/fail or all endpoints can be considered together and assessed using multivariate statistics and looking for space between site and reference samples when plotted. For invertebrate body burdens, site concentrations are compared to reference site concentrations and site concentrations are used to predict predator concentrations, which are then compared to CCME protective values. This is done on a chemical by chemical basis.

Second step is to integrate the pass/fail for the four LOEs and then compare to the rules for integrating the LOEs in Table 1 of this study. Sixteen combinations of pass/fail are possible and fall into one of four categories: 1) sediments do not present a risk,) there are adverse effects that require risk management evaluation, 3) there is a need for both risk management evaluation and further invesigation because of equivocal results and slight effects could be occurring, and 4) there is no immediate need for risk management evaluation, but further investigation is required because the impairment cannot yet be identified.

The ideas presented in this paper were developed and reviewed during a workshop funded by Environment Canada.

Grapentine, L.; Marvin, C.; Painter, S., Initial Development and Evaluation of a Sediment Quality Index for the Great Lakes Region. *Human Ecol. Risk Assess.* 2002, *8*, (7), 1549-1567.

Abstract (reproduced from article)

A sediment quality index (SQI) based on the Canadian Water Quality Index was developed and applied to the assessment of sediment quality in two Great Lakes Areas Of Concern where metals are the primary contaminants of potential concern, Peninsula Harbour (Lake Superior) and Collingwood Harbour (Lake Huron). The SQI was calculated according to an equation incorporating two elements; scope —the number of variables that do not meet guideline objectives; and, amplitude — the magnitude by which variables exceed guideline objectives. Categorizations of sediment quality were developed based on SQI scores. The robustness of the SQI was evaluated through comparison of the relative rankings of sediment quality in the two test areas with results obtained from principle components analysis (PCA) incorporating reference sites, and calculations of hazard quotients (HQs). Trends and rankings in sediment quality determined by the SQI were similar to those calculated using PCA at both test areas. The HQs also appeared to be good indicators of sediment quality. Both the SQI and HQ methods are based on existing Sediment Quality Guidelines, but the SQI had the added benefit of allowing straightforward integration of multiple contaminants. The SQI and PCA analyses appeared complementary in that the SQI incorporated information on the number of variables exceeding

guideline values and the degree to which these guidelines were exceeded. The PCA allowed a simple check of the SQI by relating test conditions to regional background. It is recommended that this analysis be performed concurrently with SQI to ensure that non-anthropogenic sources of contaminants (metals in this case) are not considered as representing an anthropogenic hazard.

Notes

WOE involved two parts: synthesizing the data within each LOE and combing the conclusions from multiple LOE to determine overall status of a site. This paper addresses integration of data within sediment chemistry LOEs. Develops sediment quality index (SQI) based on scope - the number of variables that do not meet objectives, frequency - the number of individual measurements for which objectives are not met; and amplitude - the magnitude by which variable exceed their respective objectives. Developed classifications of SQI ranging from Poor (SQI value of 0 to 44) to Excellent (SQI value of 95 to 100). Also used a hazard quotient approach and conducted principal component analysis (PCA) using large number of reference sites. The degree to which a test site falls outside the range of natural variability defined by the PCA for the reference sites is a measure of the amount of contamination.

Trends in sediment quality determined by SQI were similar to those calculated using PCA at two example test sites. For assessments with only one contaminant of concern, results suggest that the SQI and hazard quotient approaches could be undersensitive.

Menzie, C.; Henning, M. H.; Cura, J.; Finkelstein, K.; Gentile, J.; Maughan, J.; Mitchell, D.; Petron, S.; Potocki, B.; Svirsky, S.; Tyler, P., Special Report of the Massachusetts Weight-of-Evidence Workgroup: A Weight-of-Evidence Approach for Evaluating Ecological Risks. *Human Ecol. Risk Assess.* 1996, *2*, (2), 277-304.

Abstract (reproduced from article)

Weight-of-evidence is the process by which multiple measurement endpoints are related to an assessment endpoint to evaluate whether significant risk of harm is posed to the environment. In this paper, a methodology is offered for reconciling or balancing multiple lines of evidence pertaining to an assessment endpoint.

Weight-of-evidence is reflected in three characteristics of measurement endpoints: a) the weight assigned to each measurement endpoint, b) the magnitude of response observed in the measurement endpoint, and c) the concurrence among outcomes of multiple measurement endpoints. First, weights are assigned to measurement endpoints based on attributes related to: a) strength of association between assessment and measurement endpoints, b) data quality, and c) study design and execution. Second, the magnitude of response in the measurement endpoint is evaluated with respect to whether the measurement endpoint indicates the presence or absence of harm; as well as the magnitude. Third, concurrence among measurement endpoints is evaluated by plotting the findings of the two preceding steps on a matrix for each measurement

endpoint evaluated. The matrix allows easy visual examination of agreements or divergences among measurement endpoints, facilitating interpretation of the collection of measurement endpoints with respect to the assessment endpoint. A qualitative adaptation of the weight-of-evidence approach is also presented.

Notes

Professional judgment of LOEs may use both knowledge about the strengths and weaknesses of various measurements and beliefs about whether the measurements in question are likely to over- or underestimate risk. The regulator may be skeptical about the reliability of certain LOEs (e.g., field studies, as they may not have sufficient power to detect effects), whereas the risk assessor representing the regulated community may be have less confidence in LOEs that are less site-specific (e.g., comparing concentrations to literature-based benchmarks). A formal WOE could increase the risk assessor's awareness of his/her beliefs and make that more transparent to the reader.

A summary of the approach follows.

STEP 1. Eleven attributes used to select optimal measures of effects and determine confidence in line of evidence (i.e., the weight)

1. Biological linkage between measurement endpoint and assessment endpoint (or degree of association)
2. Site specificity
3. Stressor-specificity
4. Extent to which data quality objectives are met
5. Availability of an objective measure for judging environmental harm
6. Sensitivity of the measurement endpoint for detecting changes
7. Spatial representativeness
8. Temporal representativeness
9. Quantitative (can numbers be used to describe magnitude of response of ME to stressor, results are quantitative and stats can be used)
10. Correlation of stressor to response
11. Use of a standard method.

This step can be done quantitatively or qualitatively (MADEP recommends qualitatively to allow for flexibility. However, best professional judgment is used here and it must be fully documented so that the process can be transparent).

- **Quantitatively**—The workgroup determined a weight scale representing the relative importance of the attributes. The weight scale is set. Then each attribute is scored based on the definitions of scores 1 through 5. The weight of the LOE was determined by summing the products of scaling values and the scores and dividing by five.

- **Qualitatively**—LOEs are assigned a score of high, medium, or low for each of the 11 attributes. Based on those scores and on the relative importance of individual attributes, the risk assessor should determine an overall score of high, medium, or low, indicating how well the LOE represents the assessment endpoint. Risk assessors could assume all attributes are of equal importance. It is important for the risk assessor to explain their rationale for selecting the score if they are not going to use the quantitative method.

STEP 2. Determine the magnitude of response/harm: Yes/High, Yes/Low, Undetermined, No/Low, No/High.

STEP 3. Place results on a matrix to visually determine concurrence among LOEs, and see how the LOEs converge or don't converge on the matrix.

Reynoldson, T. B.; Smith, E. P.; Bailer, A. J., A Comparison of Three Weight-of-Evidence Approaches for Integrating Sediment Contamination Data within and Across Lines of Evidence. *Human Ecol. Risk Assess.* 2002, *8*, (7), 1613-1624.

Abstract (reproduced from article)

Multiple lines of evidence (LOE) are often considered when examining the potential impact of contaminated sediment. Three strategies are explored for combining information within and/or among different LOEs. One technique uses a multivariate strategy for clustering sites into groups of similar impact. A second method employs meta-analysis to pool empirically derived *P*-values. The third method uses a quantitative estimation of probability derived from odds ratios. These three strategies are compared with respect to a set of data describing reference conditions and a contaminated area in the Great Lakes. Common themes in these three strategies include the critical issue of defining an appropriate set of reference/control conditions, the definition of impact as a significant departure from the normal variation observed in the reference conditions, and the use of distance from the reference distribution to define any of the effect measures. Reasons for differences in results between the three approaches are explored and strategies for improving the approaches are suggested.

Notes

The first approach (Reynoldson et al. 2000) uses a database of reference site data. The benthic community and sediment toxicity LOEs are assessed using multivariate methods, and the chemistry is assessed using Sediment Quality Index. Each LOE is ranked 1 (excellent) to 4

(poor). The second approach (Bailer et al. 2002) quantifies a potential decrement in response at a contaminated site relative to reference sites using simple summary measures and pooled *P*-values. The third approach (Smith et al. 2002) uses a model to calcuate the probability that a site is impaired relative to reference sites for each LOE. Reference data consists of 252 sites for community data and 105 sites for toxicity data for the Great Lakes. The site data are compared to these reference data. These approaches do not account for inherent strength and limitations of each LOE. The three WOE approaches did not reach the same decisions. While each method has attractive features, the authors do not believe that one applied empirical analysis would suffice for distinguishing between alternative strategies.

These methods are quite statistical and it could be difficult to explain the methods to non-scientists. In addition, we do not often have the availability of many reference site data sets.

Reynoldson, T. B.; Thompson, S. P.; Milani, D., Integrating Multiple Toxicological Endpoints in a Decision-Making Framework for Contaminated Sediments. *Human Ecol. Risk Assess.* 2002, *8*, (7), 1569-1584.

Abstract (reproduced from article)

Contaminated sediment has been identified as one of the major impediments to ecosystem restoration, but there has been little progress made in the management of sediment contaminants. Four primary lines of evidence are generally required for informed assessments yet the integration of these various lines of evidence is problematic. Using data from 220 reference sites located in the nearshore zone of the Laurentian Great Lakes the normal response of four species of laboratory organisms to sediments representing a wide range of sediment characteristics was examined. The toxicity data from the reference sites were used to establish categories of responses to test sediments. The delineations for the three categories were developed from the standard statistical parameters of population mean and standard deviation (mean ± SD) of an endpoint measured in all reference sediments. Three approaches for integrating information were examined; the first two are score based, the third approach uses a multivariate statistical method to integrate the responses. The methods were examined using both artificial and real test site data and from this it was concluded that ordination is the superior of the three. It is the least subjective within the context of the integration of the endpoints, is quantitative, and also provides appropriate weighting based on the variation observed within reference sites.

Notes

Generally, results of toxicity tests for reference site sediments were not correlated with any specific sediment characteristic, although some trends were noted (e.g., growth and percent silt or organic carbon). Reference site data were used to establish three categories: nontoxic, potentially toxic, and toxic, based on differences from the mean (e.g., the threshold for nontoxic was set at two standard deviations below the mean). Examined three approaches for integrating the various LOE: two score-based methods and a multivariate method (ordination). Concluded

that ordination was the best method. Notes, however, that the method requires a sufficient availabiltiy of reference sites, with a minimum of 10 reference sites to one test site. The method addressed issue of subjectivity by setting effect levels transparently using *a priori* effect sizes (e.g., 2 SD or 95% probability ellipses).

Smith, E. P.; Lipkovich, I.; Ye, K., Weight-of-Evidence (WOE): Quantitative Estimation of Probability of Impairment for Individual and Multiple Lines of Evidence. *Human Ecol. Risk Assess.* 2002, *8*, (7), 1585-1596.

Abstract (reproduced from article)

Environmental decision-making is complex and often based on multiple lines of evidence. Integrating the information from these multiple lines of evidence is rarely a simple process. We present a quantitative approach to the combination of multiple lines of evidence through calculation of weight-of-evidence, with reference conditions used to define a not impaired state. The approach is risk-based with measurement of risk computed as the probability of impairment. When data on reference conditions are available, there are a variety of methods for calculating this probability. Statistical theory and the use of odds ratios provide a method for combining the measures of risk from the different lines of evidence. The approach is illustrated using data from the Great Lakes to predict the risk at potentially contaminated sites.

Notes

This paper describes a statistical approach to integrate LOEs in which the likelihood of the data is calculated under two different scenarios and a decision made based on the ratio of the likelihoods. In this approach, there are two states (the site is impaired or the site is not impaired) and we must decide which state is more likely given the data. A Bayesian approach is used – we have opinions of the site without seeing it based on previous data. After data collection, we process the data and update our opinion. A model is needed to describe the data for this statistical WOE. This approach doesn't account for inherent strength and limitations of each LOE.

These methods are quite statistical and it could be difficult to explain the methods to non-scientists. In addition, we do not often have the availability of many reference site data sets.

Sustainable Fisheries Foundation, Workshop to Support the Development of Guidance on the Assessment of Contaminated Sediments in British Columbia: Workshop Summary Report, Prepared for B.C. Ministry of Environment, Land Remediation Section; December, 2007.

Executive Summary (selected excerpts reproduced from report)

On September 24–26, 2007, the Sustainable Fisheries Foundation (on behalf of the B.C. Ministry of the Environment) convened a Workshop to Support the Development of Guidance on the Assessment of Contaminated Sediments in British Columbia. Workshop participants were challenged with the task of developing recommendations on:

- The selection of…whole-sediment and pore-water toxicity tests for evaluating risks to aquatic receptors associated with exposure to contaminated sediments
- The interpretation of the whole-sediment and pore-water toxicity tests for evaluating risks to aquatic receptors associated with exposure to contaminated sediments
- The integration of information on multiple endpoints and multiple lines-of-evidence (LOEs) to obtain a weight-of-evidence (WOE) for assessing risks to aquatic receptors associated with exposure to contaminated sediments.

Relative to the selection of toxicity tests (Work Group Session 1), workshop participants recognized that a tiered-assessment framework is used to evaluate contaminated sediments in British Columbia.… In general, it was generally recognized that the weight-of-evidence (WOE) considered should reflect the weight of the decision at sites with contaminated sediments in the province….

Workshop participants generally agreed that a suite of whole-sediment toxicity tests should be applied to assess contaminated sediments in British Columbia…

All of the work groups recognized that the results of individual toxicity tests may be used within a weight-of evidence (WOE) framework for evaluating risks to the benthic invertebrate community associated with exposure to contaminated sediments.... Workshop participants also generally agreed that such WOE evaluations require information on the magnitude of toxicity in addition to, or instead of, toxicity designation information. Hence, it was generally agreed that the information on the magnitude of the response be retained to support further analyses of the toxicity data (i.e., WOE evaluations). T he multiple category approach was considered to be useful in this respect. While WOE approaches can be defined in various ways, workshop participants generally agreed that a WOE approach is: *A tool or mechanism to improve understanding of, interpretation of, and inferences to be drawn from multiple LOEs to inform recommendations to be made by risk assessors to risk managers and site managers.* Such WOE assessments facilitate prioritization of concerns relative to the risks posed by contaminated sediments and improve the confidence that can be placed in decisions regarding the management of contaminated sediments. By integrating information from multiple LOEs to

assess risks to ecological receptors, WOE assessments provide a basis for identifying key stressors at a site, determining if something needs to be done to manage contaminated sediments, and, if so, where such remedial activities should be focused. Workshop participants identified a number of approaches that could be used to integrate multiple LOEs to assess risks to benthic invertebrates…

…[I]t was generally agreed that BCMOE should not establish prescriptive guidance on the selection of WOE approaches. Rather, practitioners should be afforded the flexibility to select the WOE approach that is most appropriate for integrating the types of data and information that were collected at a site. In addition, workshop participants developed a series of guiding principles that should be used to identify the most appropriate methods for integrating multiple LOEs at sites with contaminated sediments in British Columbia. The results of the work group discussions on all three of the topics addressed during the workshop are summarized in this document…

Notes (from Appendices):

Data Quality Assessment (comparison to Data Quality Objectives) and Data Sufficiency Assessment (power analysis) must be done prior to developing WOE approach.

Perhaps should consider equal weighting intially, with subsequent weighting of LOEs if it improves interpretation of data.

Causality is the purpose of a risk assessment. Criteria for a WOE decided during problem formulation stage, but guidance for this should not be prescriptive.

Selected Literature Reviewed in Linkov et al. 2009

Alden III, R. W.; Hall Jr., L. W.; Dauer, D. M.; Burton, D. T., An Integrated Case Study for Evaluating the Impacts of an Oil Refinery Effluent on Aquatic Biota in the Delaware River: Integration and Analysis of Study Components. *Human Ecol. Risk Assess.* 2005, *11*, (4), 879–936.

Abstract (reproduced from article)

A series of statistical and graphical techniques incorporating a "weight of evidence" approach were used to interpret results from an integrated Triad case study designed to determine potential environmental impacts to aquatic biota in the Delaware River that may be linked to PAHs found in Motiva's oil refinery effluent. Sediment concentrations of various metals, PCBs and LMW PAHs exceeding both ERL and ERM sediment quality guidelines (SQGs) were reported in the study area. However, most chemical contaminants did not exceed their respective SQGs. Results from a long-term sediment coring study indicated that there was no evidence of significant historical PAH contamination of sediments related to Motiva's exceedances. PAHs comprising the Motiva "fingerprint" were found in the surficial sediments at four near-field sites

but non-Motiva PAH concentrations (background) were shown to be significantly higher at other far-field sites (non-Motiva influence). Chronic sediment toxicity appears to have significant relationships to the patterns of most PAH isomers, certain PCB isomers, and certain metals. However, sediment toxicity does not appear to be related to the PAH isomers that are characteristic of Motiva's effluent nor to the near-field sites. Impacted benthic communities were reported in the study area, primarily at one near-field and two far-field sites. However, there were no apparent relationships between benthic community health and sediment contaminants. The status of benthic communities does not appear to be related to PAHs derived from the Motiva effluent. The "weight of evidence" analysis developed from a systematic and comprehensive series of statistical and graphical assessments indicates that, although the study area displayed some degree of sediment contamination, chronic sediment toxicity, and benthic health impacts, these environmental effects generally could not be related to Motiva's exceedances.

Notes

The authors integrated results from several previous, published studies (including Triad studies on sediment contaminants, sediment toxicity, and benthic biological effects; a sediment core study; and a PAH fingerprinting study to characterize the Motiva refinery source) to determine whether environmental effects were related to Motiva's permit exceedances for oil and grease in their effluent to the Delaware River. The authors presented a series of key questions, with separate methods, discussion, and conclusion presented for each question. Several of the methods were statistical and/or graphical in nature. The authors note that the combined results of answering each of the key questions allowed the determination of the weight of evidence for assessment of ecological effects related to Motiva effluent exceedances. Further, some individual questions relied on multiple data sets or integrative analyses and thus may represent WOE approaches. For example, the assessment of toxicity of the surficial sediments (question 3a) involved multiple endpoints standardized according to methods for plotting sediment quality triad results, followed by statistical approaches (univariate and multivariate) to assess relationships between contaminants and toxicity.

The questions and answers were reported as follows; they are presented here as they provide context for the authors' understanding and approach to WOE.

"1. Are there sediment contaminant concentrations of potential ecological significance found anywhere in the study area?

Yes, there were sediment concentrations of certain metals (primarily zinc), certain PCBs (primarily at Site DR53) and LMW PAHs that exceeded SQGs in the study area. On the other hand, most metals, individual PAH isomers, total PAHs, and chlorinated pesticides did not exceed their respective SQGs.

2. Are the patterns of sediment contaminants inferentially related to Motiva exceedances?

2.a.1. Is there evidence of significant historical contamination of sediments related to Motiva exceedances?

No, the coring study indicated that there was no evidence of significant historical contamination of sediments related to Motiva's exceedances.

2.a.2. Was there significant duration and geographic extent of historic contamination related to Motiva exceedances?

No, the coring study indicated that there was no evidence for Motiva exceedances of significant duration or geographic extent.

2.b Is there evidence of significant current (surficial layer) contamination of sediments related to Motiva exceedances?

No, there is no suggestion of significant geographic contaminant patterns in the surficial sediments that could be related to PAH exceedances from Motiva. Although the PAHs comprising the Motiva "fingerprint" were found in the sediments at the four sites in the vicinity of the discharge canal (DR1, DR2, DR23, and DR26), these sites did not display significantly elevated concentrations of most PAH isomers or concentrations of other contaminants that would be predicted to be at toxic levels. In fact, total PAH concentrations were shown to be significantly higher at other sites, particularly those that were farther downstream.

3. Are contaminant-associated biological impacts indicated? Do the following biological responses appear to be associated with concentrations of sediment contaminants?

3a. Chronic Sediment Toxicity:

3a.1. Are surficial sediments toxic in chronic toxicity tests in the laboratory?

Yes, there was some degree of chronic sediment toxicity displayed in the study area, primarily at sites DR53, DR67, DR68, and DR83.

3a.2. Is toxicity related to sediment contaminants?

Yes, there were some relationships between chronic toxicity and sediment contaminants. Chronic sediment toxicity appears to have significant relationships to the patterns of most PAH isomers, certain PCB isomers (PCB195, PCB206, and PCB209), and certain metals (particularly zinc and, to a lesser extent, mercury, but possibly also copper, arsenic, and lead). Sediment toxicity does not appear to be related to the PAH isomers that are characteristic of Motiva's effluent, most of the PCB isomers (that comprise most of the total PCB SCI values), chlorinated pesticides, nor certain metals (cadmium, chromium, and nickel).

3b. Benthic Biological Community Impacts:

3.b.1. Are benthic biological communities impacted?

Yes, there were impacted benthic communities (as indicated by MAIA IBI scores) in the study area, primarily at sites DR1, DR67, and DR68.

3.b.2. Are benthic biological impacts related to sediment contaminants?

No, there were no apparent relationships between benthic community health, as indicated by the MAIA IBI, and sediment contaminants.

4a. Are biological effects correlated with sediment contamination related to Motiva exceedances?

No, biological effects do not appear to be related to Motiva exceedances. Sites identified as being influenced by Motiva effluents did not display elevated sediment contamination, toxicity, or benthic biological impacts compared to reference sites. On the other hand, sites that are farther down the River (the Site Group Other, primarily DR53, DR67, DR68, and DR83) did display elevated indicators of sediment contamination and toxicity, but the index of benthic health indicated little impact at these sites.

4b. If biological effects are correlated with sediment contamination associated with Motiva exceedances, what is the severity and extent of these effects?

The severity of effects associated with Motiva appears to be negligible. Health of benthic communities within the study area does not appear to be greatly influenced by either the existing sediment contamination or the apparent sediment toxicity."

Burkhardt-Holm, P.; Scheurer, K., Application of the Weight-of-Evidence Approach to Assess the Decline of Brown Trout (*Salmo trutta*) in Swiss Rivers. *Aquat. Sci.* 2007, *69*, 51–70.

Abstract (reproduced from article)

To assess potential causes for the decline in catch of brown trout and their impaired health status in Switzerland, a 5-year multidisciplinary research project was conducted. Multiple causal hypotheses were postulated and investigated in a variety of laboratory and field studies. We present here the application of a weight-of evidence analysis to evaluate the results of these studies and to assess the causes for decline in brown trout abundance. Based on human health epidemiological criteria, the method considers the exposure situation, the correlation between causes and effects, specificity of effects, and amelioration due to removal. For our evaluation, we concentrated on four test rivers and included data on fish health and population density, water quality, and habitat parameters. Our results showed that proliferative kidney disease (PKD) caused by a parasite and clinical outbreak supported by other factors is a very probable single parameter for the decline of brown trout abundance at the sites of the test rivers where it occurs. Elevated levels of nitrogen compounds may also be posing a serious risk at several sites, in particular those downstream of sewage treatment plants. Several habitat parameters, such as large width, low percentage of riffles or elevated winter temperatures, were identified as factors likely contributing to impaired health, recruitment, and abundance at single sites. At most sites, more than one factor must be acting jointly to cause the observed decline in brown trout abundance.

Notes

This WOE approach is a semi-quantitative method, based on epidemiological criteria and cites the method developed by Forbes and Calow (2002). It is referred to as a retrospective ecological risk assessment or ecoepidemiology. The approach includes seven questions and an assessment of the likelihood of the potential causal factors. The method is case-specific as it relates to a particular fish population issue but can be adapted to other problems.

1. Does the proposed causal relationship make sense logically and scientifically?
2. Is there evidence of exposure to causal factor?
3. Is there evidence for association between adverse effects and presence of the causal factor, either space or time?
4. Do the exposure levels exceed quality criteria or biological thresholds?
5. Is there an effect in the population know to be specifically caused by exposure to the stressor?
6. Have results from controlled experiments in the field or laboratory led to similar effects?
7. Has removal of the stressor led to an amelioration of effects in the population?

Questions 2, 3, and 4 are addressed through site specific studies and questions 1, 5, 6, and 7 are addressed by availability of other investigations.

In this approach, each plausible parameter is stated and evaluated by the above seven questions separately. Some of the parameters in this case of reduced fish populations were considered primary because they are closely associated with anthopogenic impacts (e.g., chemical inputs, low food availability), while others are considered intermediate because they cannot be controlled and are the effects of the primary parameters (e.g., impaired health). Then the data for each parameter for each location are evaluated. The authors linked the primary causes to the adverse, intermediate effect to apply the WOE procedure. This is more of a causal analysis approach to WOE. There was no formal process for assigning weights to various paramters based on strength of association, data quality, etc.

Chapman, P. M. and Anderson, J., A Decision-making Framework for Sediment Contamination. *Integr. Environ. Assess. Manag.* 2005, *1*, (3), 163–173.

Abstract (reproduced from article)

A decision-making framework for determining whether or not contaminated sediments are polluted is described. This framework is intended to be sufficiently prescriptive to standardize the decision-making process but without using ‘‘cook book’’ assessments. It emphasizes 4 guidance ‘‘rules’’: (1) sediment chemistry data are only to be used alone for remediation decisions when the costs of further investigation outweigh the costs of remediation and there is agreement among all stakeholders to act; (2) remediation decisions are based primarily on biology; (3) lines of evidence (LOE), such as laboratory toxicity tests and models that contradict the results of properly conducted field surveys, are assumed incorrect; and (4) if the impacts of a remedial alternative will cause more environmental harm than good, then it should not be implemented. Sediments with contaminant concentrations below sediment quality guidelines (SQGs) that predict toxicity to less than 5% of sediment-dwelling infauna and that contain no quantifiable concentrations of substances capable of biomagnifying are excluded from further consideration, as are sediments that do not meet these criteria but have contaminant concentrations equal to or below reference concentrations. Biomagnification potential is initially addressed by conservative (worst case) modeling based on benthos and sediments and, subsequently, by additional food chain data and more realistic assumptions. Toxicity (acute and chronic) and alterations to resident communities are addressed by, respectively, laboratory studies and field observations. The integrative decision point for sediments is a weight of evidence (WOE) matrix combining up to 4 main LOE: chemistry, toxicity, community alteration, and biomagnification potential. Of 16 possible WOE scenarios, 6 result in definite decisions, and 10 require additional assessment. Typically, this framework will be applied to surficial sediments. The possibility that deeper sediments may be uncovered as a result of natural or other processes must also be investigated and may require similar assessment.

Notes

This study provides a framework for making decisions regarding contaminated sediments. It uses the sediment triad approach plus assessment of biomagnification potential. It is sufficiently rigid to ensure consistency between different sediment assessments. The framework can be applied to large and small sites. There are four basic rules:

1. Sediment chemistry data, such as sediment quality guidelines, will not be used alone for remediation decisions except for simple contamination problems
2. Any remediation decisiosn will be based primarily on biological responses
3. LOEs that contradicts the results of sufficiently robust field surveys are incorrect
4. A remedy will not be implemented if the remediation causes more harm than leaving the contamination in place.

The framework is tiered but more than one step can be done simultaneously.

Step 1 – Examine data

Step 2 – Develop, implement SAP, assess COCs

Step 3 – Compare to reference conditions

Step 4 –Model biomagnification potential

Step 5 – Assess sediment toxicity

Step 6 – Assess benthic community structure

Step 7 – Construct decision matrix

Step 8 – Collect additional information if needed

Step 9 – Assess deeper sediments.

LOEs are placed in a category denoting significant effect/minor or possible effect/no effect according to results. Sixteen combinations of LOE categorizations are possible and a decision matrix determines when further assessment is needed or if a management action is required.

Chapman, P. M., The Sediment Quality Triad: Then, Now and Tomorrow. *Int. J. Environ. Pollut.* 2000, *13*, (1–6).

Abstract (reproduced from article)

The past, present and future status of the Sediment Quality Triad (SQT) concept is reviewed. The SQT has developed since its inception; some early data interpretation methods remain useful and have been improved (e.g. normalising to reference data), others have not proven to be as useful as originally anticipated (e.g. a single index coupled with triangular graphical plots). SQT studies have extended to Antarctica, and the SQT concept coupled with weight of evidence forms a major basis for sediment ecological risk assessment. Future trends in the usage and utility of the SQT are suggested.

Notes

Short paper that acknowledges that the concept of reducing each component of the triad into a single index results in substantial loss of information. Recommends use of univariate and multivariate analyses, coupled with tabular decision matrices.

Chapman, P. M., Determining when contamination is pollution — Weight of evidence determinations for sediments and effluents. *Environ. Int.* 2007, *33*, 492–501.

Abstract (reproduced from article)

Contamination is simply the presence of a substance where it should not be or at concentrations above background. Pollution is contamination that results in or can result in adverse biological effects to resident communities. All pollutants are contaminants, but not all contaminants are pollutants. Differentiating pollution from contamination cannot be done solely on the basis of chemical analyses because such analyses provide no information on bioavailability or on toxicity. Effects-based measures such as laboratory or field toxicity tests and measures of the status of resident, exposed communities provide key information, but cannot be used independently to determine pollution status. Laboratory studies can be predictive, but are rarely realistic. Measures of resident communities include innate natural variability and cannot easily distinguish between adaptation to contamination (a genetic process) and acclimation (a physiological process that may decrease energy reserves, possibly reducing such critical population-level parameters as reproduction). Finally, contaminant effects may not only be direct but also indirect; predicting such effects requires knowledge of the system under study as well as appropriate use of lines of evidence (LOE) such as toxicity tests directed to key species. Consequently, in sediments, effluents or other inputs/environmental compartments, determining when contamination is or may in future become pollution, requires a weight of evidence (WOE) assessment using different LOE appropriate to the situation under investigation. WOE investigations provide two different types of information: definitive conclusions regarding pollution; or, information as to what additional, investigative studies are necessary for definitive conclusions. Effectively, a WOE assessment comprises an initial screening-level ecological risk assessment (ERA), which may be followed by a detailed-level ERA if key uncertainties need to be resolved.

Notes

Defines two types of information provided by WOE investigations: definitive conclusions regarding pollution; or information as to what additional investigative studies are necessary for definitive conclusions. Also notes that information on sediment stability should be obtained to determine whether investigations can be restricted to surficial sediments or whether they also need to consider deeper sediments. Results of LOE are summarized in a decision matrix. Toxicity and benthic community structure are given higher weight than sediment chemistry. Risks are characterized as negligible (similar to reference conditions), moderate – minor or potential differences compared to reference conditions; and high – major or significant differences compared to reference conditions. Provides some guidance and rationale for defining differences. For example, states that sediment toxicity tests are not considered different from reference unless there is greater than a 20% difference and the difference is statistically significant (Chapman and Anderson 2005).

Chapman, P. M., Traditional Ecological Knowledge (TEK) and Scientific Weight of Evidence Determinations. *Mar. Pollut. Bull.* 2007, *54*, 1839–1840.

Selected excerpts (reproduced from article)

The term ‘‘Traditional Ecological Knowledge’’ (TEK), which first came into widespread usage in the 1980s, refers to local experience acquired over long time periods of direct human contact with the environment. TEK is typically ascribed to aboriginal people who have spent their lives out on the land/waters, and who have developed a holistic understanding of the land/waters, their biota, and human interrelationships with both….

The purpose of this editorial is threefold. First, I attempt to better acquaint environmental scientists with TEK (what it is and what it is not), and provide TEK practitioners with similar information regarding scientific investigations. Second, I examine how TEK and environmental science should interact, and how they should not. Finally, I suggest that TEK not only fits with, but in fact provides a complement for scientific weight of evidence (WOE) determinations.

…the use of WOE takes scientific analytical and reductionist approaches to a higher level – a holistic level similar to TEK. For example, the concept of valued ecosystem components (VECs) and their interrelationships (i.e., knowledge of the environment as a whole) is common to both integrative scientific studies such as risk assessments and to TEK.

TEK should not be integrated into science nor should the reverse occur. As Knudtson and Suzuki (1992) note, TEK is complementary to western science, not a replacement for it. Both have value in their own right and need to be recognized as such. Resource management in particular would be best guided by a combination, not an integration of TEK and environmental science, providing an overall WOE determination in which each has equal weight. Similar separate consideration of TEK and science to develop an equally weighted WOE assessment is appropriate for a wide variety of other environmental applications, not just in North America but also in other areas of the world where indigenous people still live off the lands and waters.

Notes

This paper compares TEK and WOE. TEK is qualitative, intuitive, holistic, oral, assumes humans are part of the environment, and data are generated by resource users. WOE is quantitative (e.g., statistical), analytical, reductionist, written, assumes humans are distinct from the environment and data are generated by specialists. Both TEK and WOE are dynamic and evolve over time. The author believes that they both have their value and that TEK as well as scientific measurements could be used to assess a wide variety of environmental problems where indigenous people still live off the environment.

Critto, A.; Torresan, S.; Semenzin, E.; Giove, S.; Mesman, M.; Schouten, A. J.; Rutgers, M.; Marcomini, A., Development of a Site-Specific Ecological Risk Assessment for Contaminated Sites: Part I. A Multi-Criteria Based System for the Selection of Ecotoxicological Tests and Ecological Observations. *Sci. Total Environ.* 2007, *379*, 16–33.

Abstract (reproduced from article)

A two module Decision Support System (DSS-ERAMANIA) was developed in order to support the site-specific Ecological Risk Assessment (ERA) for contaminated sites. Within the first module, the TRIAD and the Weight of Evidence approaches were used to develop a site-specific Ecological Risk Assessment framework including three tires [*sic*] of investigation. Selected ecological observations and ecotoxicological tests were compared according to Multi Criteria Decision Analysis (MCDA) methods and expert judgment, and the obtained ranking was used to identify a suitable set of tests, at each investigation tier, to be applied to the examined case study. A simplified application of the proposed methodology, implemented in the Module 1 of the DSS-ERAMANIA, is described and discussed.

Notes

Notes one advantage of MCDA in selection of LOE is the ability to highlight similarities or conflicts among stakeholders, which can result in a deeper understanding of the values held by others. Uses a tiered Triad approach. Examines various criteria for the selection of LOE that are suitable for the various tiers of the assessment. Uses values assigned by experts to integrate weights and calculate a score for each LOE and for each Triad tier of the assessment. Based on the score, LOE are ranked according to their suitability to be applied at the various tiers of the assessment. The results depend upon the subjectivity and expertise of the experts, but their concordance can be evaluated by the system and explained to promote discussion and consensus.

Fairbrother, A., Lines of Evidence in Wildlife Risk Assessments. *Human Ecol. Risk Assess.* 2003, *9*, (6), 1475-1491.

Abstract (reproduced from article)

Methods for assessing risk to wildlife from exposure to environmental contaminants remain highly uncertain as empirical data required for accurate estimates of exposure or determination of toxicity thresholds are lacking. Some practitioners have advocated an ecological approach (i.e., "top down") to wildlife assessments to account directly for the uncertainties inherent in aggregating direct toxicological effects to individuals when estimating population risk (i.e., "bottom up" techniques). This paper suggests a methodology for conducting wildlife risk assessments that incorporates both the "bottom up" and "top down" techniques by taking into account multiple lines of evidence that are gathered by proceeding through a tiered approach including: 1) concentration of chemicals in relation to levels reported to be harmful; 2)

bioassays or toxicity studies to define dose-response relationships; and 3) field studies of population or community responses. A step-wise process progressing through these three tiers is a cost-effective method for developing the necessary information. This method is analogous to standard epidemiological approaches. Incorporation of continued monitoring and directed field studies into risk management is suggested as a means to move forward with environmental management decisions in the face of the significant uncertainties that will continue to be associated with wildlife risk assessments.

Notes

Suggests a WOE method for wildlife risk assessment. Proposes a tiered process as most cost-effective. Tier I focuses on question of whether there is reason to believe that any of the contaminants of potential concern currently exist at levels that could be directly toxic to wildlife. The next tier delineates the potential extent over which the contaminated media could have a direct impact and provides more site-specific refinements to the exposure model. The final tier provides site-associated ecological sudies to provide further evidence concerning contaminant-related effects to ecologically relevant population or community responses. Provides a flow chart of the use of lines of evidence to rule out exposure pathways, chemicals, or species of concern and to apportion risk to remaining components. Suggests that weighting of LOEs should be done only in the upper tiers, because of highly conservative nature of soil screening values.

Forbes, V. E.; Calow, P., Systematic approach to weight of evidence in sediment quality assessment: Challenges and opportunities. *Aquat. Ecosys. Health Manag.* 2004, *7*, (3), 339–350.

Abstract (reproduced from article)

Sediments are complex systems, and contaminants interact with them in a multitude of ways that influence exposure and effects on sediment-dwelling biota. Likewise a variety of non-chemical agents may act on benthic communities; these may dominate, contribute to or mask any effects from chemical contaminants. The first problem is to recognize that adverse effects have occurred. Once effects have been convincingly identified it is then necessary to assess likely causal agents retrospectively. Here we present a series of questions, elaborated from human health epidemiology, that can help to structure retrospective sediment assessments. We propose a method that aims to guide interpretation of various combinations of answers to the questions so that conclusions about the likelihood that identified agents have caused the observed effects in sediment systems can be consistently drawn. We demonstrate the approach by applying it to two published case studies. The first deals with estuarine communities in the northern Gulf of Mexico exposed to a wide variety of contaminants originating from a range of sources. The second involves a freshwater stream community impacted by road runoff. Although simpler than other weight of evidence approaches, we believe that the method provides a systematic, explicitly documented and consistent approach that may be particularly

effective for defining priorities in situations where the evidence is limited and/or at least partly qualitative.

Notes

This paper describes a WOE approach to identify likely causes of observed adverse ecological effects in sediment quality assessment. It is referred to as either a retrospective risk assessment or ecoepidemiology. It uses human health epidemiological criteria (Hill criteria) or causality criteria. Seven quesitions specific to sediment quality are addressed through this approach:

1. Is there evidence that the sediment ecosystem has been exposed to the suspected agent(s)?
2. Are there correlations between effects and exposure in space or tiem?
3. Is there exceedance of accepted sediment quality guidelines?
4. Have the observed effects been confirmed in controlled experiments?
5. Has removal of the suspected agent(s) led to amelioration?
6. Are there agent-specific effects?
7. Are the relationships between suspected agent(s) and observed effects logical and scientific?

This WOE approach is done on a chemical by chemical basis. This approach can be used even if there are data gaps and unanswered questions. Professional judgment is usually used in coming to conclusions and it plays an important role. The approach, however, encourages documentation of all considerations before in making conclusions.

Good, I. J. Weight of evidence and the Bayesian likelihood ratio. In: Aitken C.G.G., Stoney D., editors. *The Use of Statistics in Forensics Science*. Boca Raton: CRC Press; 1991. p. 85–106.

No Abstract Provided

Notes

This paper provides a discussion of the theory of WOE from a statistical point of view. It does not provide a specific WOE approach for integrating LOEs.

Hall, L. W. J.; Dauer, D. M.; III, R. W. A.; Uhler, A. D.; DiLorenzo, J.; Burton, D. T.; Anderson, R. D., An Integrated Case Study for Evaluating the Impacts of an Oil Refinery Effluent on Aquatic Biota in the Delaware River: Sediment Quality Triad Studies. *Human Ecol. Risk Assess.* 2005, *11*, (4), 657–770.

Abstract (reproduced from article)

Triad studies consisting of chemical characterizations in sediment, sediment toxicity testing, and benthic community assessments were used to determine the impacts of Motiva Enterprises oil refinery effluent [primarily polynuclear aromatic hydrocarbons (PAHs)] on aquatic biota in the Delaware River. Triad studies were conducted at 15 near-field, mid-field, and far-field sites near the Refinery in the Delaware River during the spring and summer of 2001 and 2002. Fingerprinting analysis showed that Motiva-related PAHs may be present at four near-field sites. A summary of all Triad data by site for 2001 shows a strong case for contaminant-induced degradation at one near-field site in the discharge canal of the Refinery and two far-field sites as all three lines of evidence suggest impairment. Stressful conditions for benthic communities at the near-field site include elevated temperature conditions and various pesticides (Dieldrin, 4,4′-DDD and 4,4′-DDT). Toxicity at the near-field site may also be related to the presence of pesticides exceeding sediment quality guidelines. Due to exceedances of individual Effects Range Low (ERL) guidelines for two individual PAHs, the Motiva effluent cannot be eliminated as a potential stressor at the near-field site during the summer of 2001. A summary of Triad data for the 15 Delaware River sites sampled in 2002 shows only one mid-field site where all three lines of evidence suggest impairment. Toxicity and benthic community impairment at this mid-field site may be related to PCBs and low molecular weight PAHs. Three individual PAH ERL values were exceeded at three near-field sites in 2002. The source of these PAHs is a combination of both background signature and the Motiva effluent. Multivariate analysis, using a weight of evidence approach, is used to address ecological effects of the Motiva effluent in more detail in Alden *et al.* (2005).

Notes

Presents a summary of Sediment Quality Triad data. WOE analysis is presented in companion paper (Alden et al. 2005).

Krimsky, S., The Weight of Scientific Evidence in Policy and Law. *Am. J Publ. Health* 2005, *95*, (S1), S129-S136.

Abstract (reproduced from article)

The term “weight of evidence” (WOE) appears in regulatory rules and decisions. However, there has been little discussion about the meaning, variations of use, and epistemic significance of WOE for setting health and safety standards. This article gives an overview of the role of WOE in regulatory science, discusses alternative views about the methodology underlying the concept, and places WOE in the context of the Supreme Court’s decision in *Daubert v Merrell*

Dow Pharmaceuticals, Inc (1993). I argue that whereas the WOE approach to evaluating scientific evidence is gaining favor among regulators, its applications in judicial processes may be in conflict with some interpretations of how the Daubert criteria for judging reliable evidence should be applied.

Notes

This paper reviews the existing WOE approaches in risk assessment and discusses how they are used in policy and law. No WOE approach or framework is described. The author notes that there are formal WOEs as well as seat-of-the-pants WOEs. The author states that while WOE is gaining traction in the regulatory world, a recent court decision (Daubert v. Merrell Dow Pharmaceuticals) could set a precedent for dismissing the scientific WOE approach and even the data used in the WOE.

Linkov, I.; Loney, D.; Cormier, S.; Satterstrom, F. K.; Bridges, T., Weight-of-Evidence Evaluation in Environmental Assessment: Review of Qualitative and Quantitative Approaches. *Sci. Total Environ.* 2009, *doi:10.1016/j.scitotenv.2009.05.004*.

Abstract (reproduced from article)

Assessments of human health and ecological risk draw upon multiple types and sources of information, requiring the integration of multiple lines of evidence before conclusions may be reached. Risk assessors often make use of weight-of-evidence (WOE) approaches to perform the integration, whether integrating evidence concerning potential carcinogenicity, toxicity, and exposure from chemicals at a contaminated site, or evaluating processes concerned with habitat loss or modification when managing a natural resource. Historically, assessors have relied upon qualitative WOE approaches, such as professional judgment, or limited quantitative methods, such as direct scoring, to develop conclusions from multiple lines of evidence. Current practice often lacks transparency resulting in risk estimates lacking quantified uncertainty. This paper reviews recent applications of weight of evidence used in human health and ecological risk assessment. Applications are sorted based on whether the approach relies on qualitative and quantitative methods in order to reveal trends in the use of the term weight of evidence, especially as a means to facilitate structured and transparent development of risk conclusions from multiple lines of evidence.

Notes

Defines WOE as “a framework for synthesizing individual lines of evidence, using methods that ae either qualitative (examining distinguishing attributes) or quantitative (measuring aspects in terms of magnitude) to develop conculsions regarding questions concerned with the degree of impairment or risk. Table 1 presents the following classification of various WOE methods.

- Listing Evidence: Presentation of individual LOE without integration
- Best Professional Judgment: Qualitative integration of individual LOE
- Causal Criteria: A criteria-based method for determining cause and effect
- Logic: Standardized evaluation of individual LOE based on qualitative logic
- Scoring: Quantitative integration of multiple LOE using simple weighting or ranking
- Indexing: Integration of LOE into a single measure
- Quantification: Integrated assessment using formal decision analysis and statistical methods.

A review of 44 ecological studies found that qualitative methods are used most often, particularly best profession judgment (8 studies), causal criteria (10 studies), or logic (15 studies). States that although qualitative analysis of individual lines of evidence or even quantitative analysis using Scoring, Indexing, and Statistical methods can be useful for decision making, they do not include options for quantitatively integrating decision-maker values and judgment. The main advantage of MCDA-based WOE is its method of integrating individual LOEs, as well as for evaluating the sensitivity of the conclusions to changes in weighting and integration, which can be used for debate and to reach consensus.

Lowell, R. B.; Culp, J. M.; Dube, M. G., A Weight-of-Evidence Approach for Northern River Risk Assessment: Integrating the Effects of Multiple Stressors. *Environ. Toxicol. Chem.* 2000, *19*, (4), 1182–1190.

Abstract (reproduced from article)

Northern river ecosystems are subject to a variety of stressors having multifaceted (and sometimes opposing) effects, making interpretation at a regional scale difficult. We have addressed this problem by using a weight-of-evidence approach that combines analysis of field data (to determine patterns) with experimental hypothesis testing (to determine mechanisms). Two of the more important sources of aquatic impacts in western Canada are pulp mill and municipal effluents. Their regional impacts on benthic biota were evaluated for two major river systems, the Thompson and Athabasca rivers, using an integrative approach. In the more southerly Thompson River, several lines of evidence (including field and laboratory experiments, field sampling over a 20-year period, and isotopic analysis) led to the conclusion that, although some toxic effects were apparent, these effects were usually masked by the (sometimes excessive) nutrient enhancement effects of these effluents, sometimes via novel pathways. Furthermore, analysis of the data revealed a fairly delicate balance in effluent treatment involving trade-offs between the negative effects of toxic contaminant loading versus a switch to a more eutrophic community. In the more northerly Athabasca River, effluent effects can be modified by the added impact of another stressor: widespread winter freeze-up,

which prevents reaeration of oxygen depleted waters, coupled with low dissolved oxygen levels in the substratum where benthic invertebrates are found, resulting in a net shift in effluent effect from one of nutrient enhancement to a more inhibitory effect. Advantages to applying formalized causal criteria, as outlined in this weight-of-evidence approach, include helping to tie together diverse assemblages of data on the effects of multiple stressors and identifying important informational gaps, thus making ecological risk assessments more rigorous and robust.

Notes

This paper describes a WOE approach for integrating multiple stressors into an evaluation of ecological risk. Three steps are critical: 1) establishing causality, 2) defining acceptable limits, and 3) linking environmental components in a decision-making framework. These all fall into the later stages (analysis and risk characterization) of the U.S. EPA ERA guidance.

Establishing casuality. The epidemiological criteria are used to assess the strength of the causal link. These criteria include: 1) spatial correlation of stressor and effect along gradient from more to less exposed areas, 2) temporal correlation for stressor and effect relative to time course of exposure, 3) plausible mechanism linking stressor and effect, 4) experimental verification of stressor effects under controlled condition, 5) strength, 6) specificity, 7) evidence of exposure , 8) consistency of stressor-effect associations and 9) coherence with existing knowledge. They are based on those put forth by Fox (1991), Suter (1993), Beyers (1998) and Gilbertson (1997).

Define acceptable limits. The magnitude of the effect must be compared a critical effect that is judged to be ecologically significant.

Linking environmental components in a decision-making framework. Once causality has been established and effect exceeds a critical threshold, each component must be weighted and summed to determine if and what action is needed. The article does not describe how each component is weighted.

This paper provides an example application of this framework to assess risk to Canadian rivers from pulp mill effluent and municial effluent.

McDonald, B. G.; Chapman, P. M., Selenium Effects: A Weight-of-Evidence Approach. *Integrated Environ. Assess. Manag.* 2007, *3*, (1), 129–136.

Abstract (reproduced from article)

Selenium is increasingly an issue for a wide range of mining, industrial, and agricultural operations. Appropriate methods for evaluating the impacts of selenium in aquatic ecosystems are vigorously debated in the literature. Two common approaches include the use of tissue residue guidelines and reproductive toxicity testing using field-collected fish; however, each approach on its own does not provide sufficient evidence that wild fish populations are in fact

impaired. The limitations of each method are discussed, and recommendations to improve the relevance of each line of evidence are provided. A 3rd line of evidence, field measurement of fish population dynamics, is proposed and also discussed. A framework, consistent with an ecological risk assessment methodology, for the design, application, and interpretation of selenium weight-of evidence investigations is proposed.

Notes

This paper describes a WOE framework specifically related to evaluating the effects of selenium in fish. The framework consists of three LOEs: 1) Comparison of selenium concentrations in fish tissue to target residue guidelines to protect against adverse effects associated with selenium in aquatic systems, 2) reproductive toxicity testing using fertilized eggs from field-collected fish, and 3) assessment of fish populations at the site. The LOEs could be conducted in a tiered fashion or LOEs could be conducted simultaneously. Potential outcomes of the WOE given the results of one or all three LOEs are provided. Greater weight is given to biology data than chemistry data.

McDonald, B. G.; deBruyn, A. M.; Wernick, B. G.; Patterson, L.; Pellerin, N.; Chapman, P. M., Design and Application of a Transparent and Scalable Weight-of-Evidence Framework: An Example From Wabamun Lake, Alberta, Canada. *Integrated Environ. Assess. Manag.* 2007, *3*, (4), 476–483.

Abstract (reproduced from article)

A weight-of-evidence (WOE) framework was developed to evaluate potential effects on the aquatic ecosystem of Wabamun Lake (Alberta, Canada) associated with the release of Bunker ‘‘C’’ oil after a train derailment. The wide variety of stakeholders and interested regulatory agencies made it necessary to develop a consistent and transparent approach to assessing ecological effects on multiple ecosystem components within the lake with the use of a large number of lines of evidence (LOEs). Consequently, a scalable WOE framework was necessary to integrate the findings of 38 different LOEs. A priori and a posteriori weighting factors were applied to each individual LOE, and a combination of numeric and nonnumeric rating systems was used to integrate LOEs into an overall WOE conclusion for 5 different ecosystem components. We provide guidance regarding the development of a WOE framework and emphasize techniques that enhance the application of best professional judgment during the WOE process.

Notes

For this case study, the presence of multiple, pre-existing stressors made it necessary to assemble a large number of LOEs. Uses a hybrid of numeric and non-numeric WOE approaches. Each LOE was rated by comparison to benchmarks or reference conditions (low, moderate, severe). Symbolic ratings were temporarily converted to numbers and weighted according to quality and relevance of information, using a combination of *a priori* and *a*

posteriori weighting factors. The weighted numerical ratings for each LOE were integrated into an overall numerical WOE score and converted back into a non-numerical WOE rating.

A priori weighting factors:

- Representativeness
- Methodological robustness
- Clarity of interpretation
- Permanence of effects

A posteriori weighting factors:

- Coherence of response
- Evidence of causality

A final table related numerical scores to outcomes, including: *negligible effects, moderate effects,* and *severe effects.*

Highly transparent and relatively simple framework.

McPherson, C.; Chapman, P. M.; deBruyn, A. M. H.; Cooper, L., The Importance of Benthos in Weight of Evidence Sediment Assessments — A Case Study. *Sci. Total Environ.* 2008, *394*, 252 - 264.

Abstract (reproduced from article)

Sediment quality in a Texas reservoir subject to point and non-point sources of contaminants was assessed using the Sediment Quality Triad weight of evidence approach. Fifteen stations were sampled plus a reference station which, unfortunately, comprised a different habitat type than the other 15 stations. Accordingly, standard comparisons between reference and exposed stations were inappropriate. Interpretation of potential relationships between benthic community structure and sediment-associated contaminants was also confounded by differences in habitat-related characteristics (e.g., water depth and total organic carbon) within the reservoir. Multivariate analyses of the benthic community identified two station groupings separated primarily by habitat-related differences rather than contaminant-related toxicity. Laboratory toxicity tests and chemical analyses, including measures of bioavailability, did not differ consistently between the two community-based station groupings, indicating that toxicity resulting from chemical contamination was not the primary factor in observed community structure in the reservoir, although alterations to the benthos due to chemical contamination could not be ruled out in the absence of an appropriate reference comparison. Appropriately giving highest weight to resident benthic community structure, followed by the results of

laboratory toxicity tests, then chemical analyses, provided the best possible assessment of chemical pollution in the absence of a suitable reference comparison. The alternative approach of relying on only sediment toxicity and chemistry data, without considering the full weight of evidence, would have provided misleading information.

Notes

A WOE approach (citing Chapman papers, Menzie et al. 1996 papers) was used to assess sediment quality in a reservoir in Texas using sediment triad data and integrated into a single 'balance of probabilities' conclusion. This study was designed expecting that the reference station would be representative of background conditions and that the benthic community LOE would be assessed using the magnitude and significant differences in benthic metrics relative to that reference station. This was not the case, so the decision framework for evaluation of the benthic community LOE required *a posteriori* re-evaluation and modification. This paper describes how each LOE is intrepreted and integrated on a station by station basis. While LOEs are assigned weight in terms of confidence and the rationale for the weight is given, there was not a formal approach to this outlined in the paper as there is in Menzie et al. (1996). Sediment chemistry was given minimal weight, benthic data were given high weight, and sediment toxicity was given medium weight. Casuality was assessed through correlation analyses.

Moraes, R.; Gerhard, P.; Andersson, L.; Sturve, J.; Rauch, S.; Molander, S., Establishing Causality between Exposure to Metals and Effects on Fish. *Human Ecol. Risk Assess.* 2003, *9*, (1), 149-169.

Abstract (reproduced from article)

This study evaluates causal relationships between chronic exposure of fish to metals and effects at different levels of biological organization based on a weight-of evidence approach. Criteria for evaluation of causality were strength, consistency, and specificity of the association, as well as biological gradient and plausibility. Field sampling was conducted three times between 1998 and 2000, in Furnas Stream, impacted by an abandoned lead mine, and in three other locations, including two reference and one impacted sites. Levels of Pb, Zn, Cd, and Ag in sediments from the Furnas Stream exceeded background levels, and their concentrations were above sediment quality guidelines. Residual levels of metals in fish tissue were high enough to indicate reduced growth, reproduction and/or survival according to toxicological benchmarks. Lead-induced biochemical changes (ALA-D activity depletion) were observed in two species of siluriform catfish. The condition factor of a predatory catfish was reduced, and the percentage of prey generalists was higher in Furnas than at the noncontaminated sites. Reduction in fish community diversity and density was observed. Integration of data provided supporting evidence that observed effects on fish from the Furnas Stream resulted from long-term exposure to metals, however influences from other stressors cannot be ruled out.

Notes

This paper describes a causally-based WOE approach to evaluating potential effects of heavy metals in fish. The LOEs (e.g., body burden, enzyme activity, condition factor, community diversity) reflected many levels of biological organization and varying degress of specificity. The first step is to develop a conceptual model linking stressors to receptors. The second step is to select LOEs based on the conceptual model that could indicate causality between exposure and effect. The third step is to design and peform the biological surveys in contaminated and non-contaminated areas. A key to this evaluation is the selection of appropriate reference areas. Then, the LOEs are combined using Suter et al. (2000) WOE approach. It relies on establishing causal associations between stressors and effects on biota. For each LOE, the following criteria were considered:

1. Strength of association
2. Consistency of the association
3. Specificity of the association
4. Biological gradient
5. Biological plausibility.

The authors developed a matrix that summarized the causal criteria applied to the different LOEs. The criteria either supported the LOE, did not support the LOE, evidence was too ambiguous to interpret, or the criterion was not applicable to that LOE.

The specificity criterion was not met in the majority of cases, because the effect could have been caused by anthropogenic or natural causes.

Also, because of time and resources restrictions, it is often difficult to fully evaluate spatial and temporal variability. Therefore, heavy reliance is placed on comparison of site conditions to reference site conditions.

Semenzin, E.; Critto, A.; Carlon, C.; Rutgers, M.; Marcomini, A., Development of a Site-Specific Ecological Risk Assessment for Contaminated Sites: Part II. A Multi-Criteria Based System for the Selection of Bioavailability Assessment Tools. *Sci. Total Environ.* 2007, *379*, 34–45.

Abstract (reproduced from article)

A comparison procedure based on Multi-Criteria Decision Analysis (MCDA) and expert judgment was developed in order to allow the comparison of bioavailability tests to implement the chemical Line of Evidence (LOE) within a TRIAD based site-specific Ecological Risk Assessment framework including three tiers of investigation. The proposed methodology was included in the Module 1 of the Decision Support System DSS-ERAMANIA and the obtained

rank supported the selection of a suitable set of available tests to be applied to the case study. A simplified application of the proposed procedure is described and results obtained by the system software are discussed.

Notes

This paper describes a process selecting suitable bioavailability tools to be used for the chemical LOE in the sediment TRIAD at three different tiers. Tiers 1 and 2 use indirect measures of bioavailability and Tier 3 uses biota-specific direct estimates of bioavailability. The first tier uses cheap and quick tools and more costly and lengthy tools are preferred in Tiers 2 and 3. This paper has a companion paper (Critto et al. 2007), which describes the full multi-criteria decision analysis. The tools reliability is also considered. First, tier-specific weights are assigned to all the comparative criteria to specify the relevance of each criterion for each TRIAD tier. Then, numerical equivalents are assigned to the ratings of the criteria so they are all expressed similarly. Then, the numerical criteria are normalized and weights and numerical evaluations are aggregated into tier-specific scores by a linear MCDA method. A system expert and expert of TRIAD-based ERAs are needed. This model can handle heterogeneous comparative criteria and integrates the evaluation provided by different experts.

Semenzin, E.; Critto, A.; Rutgers, M.; Marcominia, A., Integration of Bioavailability, Ecology and Ecotoxicology by Three Lines of Evidence into Ecological Risk Indexes for Contaminated Soil Assessment. *Sci. Total Environ.* 2008, *389*, 71-86.

Abstract (reproduced from article)

A Weight of Evidence approach was applied to define three integrated effect indexes estimating the impairment on terrestrial ecosystems caused by the stressor(s) of concern. According to a Triad approach, the integrated effect indexes combined the information provided by the measurement endpoints of each line of evidence (chemistry/bioavailability, ecology and ecotoxicology) and allowed to analyse the impairment degree highlighted by each measurement endpoint as difference from the reference condition. Multi-Criteria Decision Analysis (MCDA) was used for the aggregation of the complementary Triad information, including expert judgment and a weighted procedure based on the endpoint sensitivity and the sensitivity of the test for ecosystem effects. The developed methodology was implemented in the DSS-ERAMANIA, Module 2, and is presented in this paper as "Integrated Effect Indexes" (IEI) sub-module. The latter has been preliminary applied to the Acna di Cengio (Italy) contaminated site; the results of this application are presented and discussed.

Notes

Discusses WOE approaches based on matrices and indices. Although use of indices has been criticized because of loss of information, use of indices can facilitate communication with stakeholders. Used Multi-Criteria Decision Analysis (MCDA) to develop a decision support tool, "Integrated Effect Indexes" (IEI) as the second module of the Decision Support System

DSS-ERAMANIA (Semenzin et al. 2007). IEI is designed to aggregate various LOEs and quantify impairment. The procedure is based on expert judgment and follows five steps:

1. Classification of measurement endpoints
2. Assignation of impairment thresholds
3. Development of normalization functions
4. Comparison with reference conditions
5. Identification of impairment classes for the endpoints.

Presents a color-coded table of values, including the impairment values for the various LOEs that were weighted by means of values assigned by experts. However, the rationale for the assignment of the values is not described.

Weed, D. L., Weight of Evidence: A Review of Concept and Methods. *Risk Anal.* 2005, *25*, (6), 1545-1557.

Abstract (reproduced from article)

"Weight of evidence" (WOE) is a common term in the published scientific and policy-making literature, most often seen in the context of risk assessment (RA). Its definition, however, is unclear. A systematic review of the scientific literature was undertaken to characterize the concept. For the years 1994 through 2004, PubMed was searched for publications in which "weight of evidence" appeared in the abstract and/or title. Of the 276 papers that met these criteria, 92 were selected for review: 71 papers published in 2003 and 2004 (WOE appeared in abstract/title) and 21 from 1994 through 2002 (WOE appeared in title). WOE has three characteristic uses in this literature: (1) metaphorical, where WOE refers to a collection of studies or to an unspecified methodological approach; (2) methodological, where WOE points to established interpretative methodologies (e.g., systematic narrative review, meta-analysis, causal criteria, and/or quality criteria for toxicological studies) or where WOE means that "all" rather than some subset of the evidence is examined, or rarely, where WOE points to methods using quantitative weights for evidence; and (3) theoretical, where WOE serves as a label for a conceptual framework. Several problems are identified: the frequent lack of definition of the term "weight of evidence," multiple uses of the term and a lack of consensus about its meaning, and the many different kinds of weights, both qualitative and quantitative, which can be used in RA. A practical recommendation emerges: the WOE concept and its associated methods should be fully described when used. A research agenda should examine the advantages of quantitative versus qualitative weighting schemes, how best to improve existing methods, and how best to combine those methods (e.g., epidemiology's causal criteria with toxicology's quality criteria).

Notes

This paper is a review of WOE used in science and policy literature from 1994 to 2002. The results of the review indicated the following:

- WOE is poorly defined. It's important for the user to be clear about its definitions, its uses and its implications.
- WOE as a metaphor is the most common use of WOE in the literature. There is no description of or reference to a formal method. This method highlights a particular problem with WOE methods: lack of transparency.
- WOE as methodology. These are studies where WOE is a methodological approach in that all evidence is examined and interpreted. Many articles in this category did not define what is meant by all evidence and no WOE method may be described.
- Familiar WOE may include systematic narrative reviews, quality criteria reviews for toxicological studies, epidemiology's causal criteria, meta-analysis, mixed epidemiology-toxicological methods, and quantitative weighting schemes.
  - Systematic narrative reviews describe the state of the science, make research recommendations, make claims about casuality, and make public health recommendations.
  - Quality criteria for toxicological studies to assign reliability categories (e.g., reliable without restriction, not reliable) to scientific studies. Unreliable data are not used in the WOE, so this approach does not use all of the the evidence and a weighting scale.
  - Epidemiology's causal criteria (Hill's criteria) reviews the body of literature using nine criteria (consistency of association, strength of association, dose response, temporality, experimentation, specificity, biological plausibility, coherence, and analogy). Some users of this methodology do not define rules for evaluating evidence.
  - Meta-analysis is often used to summarize and weight evidence from several human population studies. It provides a weighted average estimate of effect across several studies. It also provides precise estimates of the magnitude of effects and the dose-response relationships but judgment is used to address the causal relevance of these estimates.
  - Mixed epidemiology-toxicology methods, which is exemplified by a criteria-based method of causal inference similar to that of Hill but also included a number of other considerations used to judge nonhuman evidence.

- Quantitative methods include those described for WOE in ERAs of hazardous waste sites, toxicologically-based WOE for ranking chemicals on their endocrine disruption potential, and a method used to determine the extent to which chemicals have interactive effects when in mixtures.

The author notes that the main problem with WOE is that there are mutliple definitions and uses. There are also problems with different kinds of weighting schemes and how to go about weighting (e.g., quantitative or qualitative). A solution to these problems is for the user of WOE to fully describe the WOE methods that are being employed. The author states that judgment is necessary in WOE and it is important to understand how it is used in the WOE.

This article presents three options for the future of WOE: 1) encourage that WOE methods be fully described when used so that a consensus of the meaning and methods of WOE could be reached, 2) abandon WOE and develop a research agenda on familiar interpretative methods, such as causal criteria and meta-analysis, or 3) accept the variety of WOEs and encourage users to fully define the approach and the methods.

Notes

This paper is a review of WOE used in science and policy literature from 1994 to 2002. The results of the review indicated the following:

- WOE is poorly defined. It's important for the user to be clear about its definitions, its uses and its implications.
- WOE as a metaphor is the most common use of WOE in the literature. There is no description of or reference to a formal method. This method highlights a particular problem with WOE methods: lack of transparency.
- WOE as methodology. These are studies where WOE is a methodological approach in that all evidence is examined and interpreted. Many articles in this category did not define what is meant by all evidence and no WOE method may be described.
- Familiar WOE may include systematic narrative reviews, quality criteria reviews for toxicological studies, epidemiology's causal criteria, meta-analysis, mixed epidemiology-toxicological methods, and quantitative weighting schemes.
  - Systematic narrative reviews describe the state of the science, make research recommendations, make claims about casuality, and make public health recommendations.
  - Quality criteria for toxicological studies to assign reliability categories (e.g., reliable without restriction, not reliable) to scientific studies. Unreliable data are not used in the WOE, so this approach does not use all of the the evidence and a weighting scale.
  - Epidemiology's causal criteria (Hill's criteria) reviews the body of literature using nine criteria (consistency of association, strength of association, dose response, temporality, experimentation, specificity, biological plausibility, coherence, and analogy). Some users of this methodology do not define rules for evaluating evidence.
  - Meta-analysis is often used to summarize and weight evidence from several human population studies. It provides a weighted average estimate of effect across several studies. It also provides precise estimates of the magnitude of effects and the dose-response relationships but judgment is used to address the causal relevance of these estimates.
  - Mixed epidemiology-toxicology methods, which is exemplified by a criteria-based method of causal inference similar to that of Hill but also included a number of other considerations used to judge nonhuman evidence.

- Quantitative methods include those described for WOE in ERAs of hazardous waste sites, toxicologically-based WOE for ranking chemicals on their endocrine disruption potential, and a method used to determine the extent to which chemicals have interactive effects when in mixtures.

The author notes that the main problem with WOE is that there are mutliple definitions and uses. There are also problems with different kinds of weighting schemes and how to go about weighting (e.g., quantitative or qualitative). A solution to these problems is for the user of WOE to fully describe the WOE methods that are being employed. The author states that judgment is necessary in WOE and it is important to understand how it is used in the WOE.

This article presents three options for the future of WOE: 1) encourage that WOE methods be fully described when used so that a consensus of the meaning and methods of WOE could be reached, 2) abandon WOE and develop a research agenda on familiar interpretative methods, such as causal criteria and meta-analysis, or 3) accept the variety of WOEs and encourage users to fully define the approach and the methods.