FACT SHEET

09

JANUARY 2016

METALS ENVIRONMENTAL RISK ASSESSMENT GUIDANCE

MERAG

Weight of evidence approach

Check you have the most recent fact sheet by visiting www.icmm.com

ARCHE

Eurometaux
European Association of Metals

ICMM
International Council on Mining & Metals

CONTENTS

The content of the Metal Risk Assessment Guidance (MERAG) fact sheets reflect the experiences and recent progress made with environmental risk assessment methods, concepts and methodologies used in Chemicals Management programs and Environmental Quality Standards setting (soil, water, sediments, …) for metals. Because science keeps evolving, these fact sheets will be updated on a regular basis to take into account new developments.

1. INTRODUCTION

Within a risk assessment framework, quite often the precautionary principle (PP) is applied. The PP paradigm states that if an action has a suspected risk of causing harm to the public or to the environment, precautionary measures should be taken even if some cause and effect relationships are not fully established scientifically. This approach allows policymakers to make decisions in situations when extensive scientific knowledge on the matter is lacking. **When more scientific evidence is available, the weight-of-evidence (WoE) approach could help policymakers make a more balanced decision on potential harm caused by the use of a substance.**

The WoE approach is a phrase that has often been used in recent years in the field of environmental assessment, implying that multiple and differing types of data are used to support conclusions on dominant stressors, biological impairment or risk (Burton et al 2002a). The term “weight of evidence” constitutes neither a scientifically well-defined nor an agreed formalised concept characterised by defined tools and procedures (Weed 2005). A review of the literature shows that it could be quite simplistic, such as where two approaches were used and there was a best professional judgement used to decide “weight”. Conversely, advanced multivariate statistics have been used to define relationships with multiple lines of evidence (LoE values) (Burton et al 2002a). The process incorporates judgements about the quality, extent and congruence of the information in each LoE. Similar terminology has included “burden of evidence”, “strength of evidence” or “multiple lines of evidence”. Other scientists have actually “weighted” specific data, giving them more importance than other data (eg a biotic response vs a chemical concentration). This approach, however, requires subjective judgements on a weighting factor (eg 4x vs 1x). These widely varying applications of the WoE have been problematic and confusing and now create suspicion among some scientists, as the quality of the process can vary dramatically.

The WoE process can help determine the extent of pollution, its ecological significance, the optimal remedial options and the urgency of corrective actions. It is often used within the context of ecological risk assessments. Evaluations should be both observational (eg ecologically based) and investigative (eg toxicological determinations of cause and effect in relation to chemical and/or physical stressors). The term WoE suggests that a level of

certainty exists with the assessment of conclusions when, in fact, there may continue to be significant uncertainty in the conclusions. This misconception can impede the decision-making process that links the assessment with remediation alternatives, thereby resulting in incorrect management decisions that may be over- or under-protective of human health and resident biota.

There is no standardised method or regulatory guidance on how to conduct WoE studies. The Sediment Quality Triad (Long and Chapman 1985; Chapman 1990, 1996), the consensus-based approach of Menzie et al (1996), considerations recommended for relative chemical rankings (Swanson and Socha 1997) and the WoE framework (Burton et al 2002b) appear to be the only approaches published in the open literature that provide any degree of guidance on conducting environmental WoE assessments.

The WoE process combines facts with expert judgement, especially so in a qualitative WoE process. Though it is challenging, the process should include several systematic steps that are transparent and effectively communicated to interested parties (ie stakeholders). The quality of the process can be compared to the quality of a model and its predictive capability. A model that ably characterises a system and its critical interactions, with high-quality data as its foundation, and a qualified user, will provide more realistic predictions. The WoE process is such a model. A well-defined conceptual model with clearly defined technical question(s) and highly qualified scientists as the users will allow for a more meaningful WoE decision-making process (Peters et al 1998). Unfortunately, all scientists have their own unique biases and expertise limitations. For this reason, the WoE process becomes increasingly credible, as statistical certainty is brought into the process, as opposed to subjective judgements of weight.

Improvements have been made to the WoE process, such as by defining a “consensus-based” approach or through the use of a variety of quantitative methods for improved integration of multiple LoE values (Burton et al 2002a; Wiegers et al 1998). Yet, it is apparent that no single WoE study design is appropriate for all assessments of ecosystem impairment, given the wide range of stakeholder concerns and resource availability, and the differences in ecosystems, study design, expertise and execution. Nevertheless, a standard WoE framework that contains essential “certainty elements” is

useful (Burton et al 2002b). "Certainty" is used to describe the elements because its inclusion reduces uncertainty during each step of the process.

Typically, studies of ecosystem impairment have included the following LoE values:

- chemical concentrations (eg ambient concentrations, criteria/standards)
- resident biota (eg species, populations, community indices)
- laboratory-based toxicity responses of surrogate organisms
- tissue chemical residue concentrations
- model predictions of fate and/or effects (eg contaminant transport, bioaccumulation, chemical toxicity probability distributions).

On occasion, assessment approaches have included habitat suitability, *in situ*-based toxicity and bioaccumulation, behavioural responses, subcellular biomarker responses and ecosystem functioning. The first three LoE values are the most commonly used and comprise the original Sediment Quality Triad approach (Long and Chapman 1985), which has been widely used. Rarely in these multiple LoE assessments have the data generated from each LoE been linked and analysed statistically, thereby providing a strong WoE-based conclusion (Burton et al 2002a, 2002b).

1.1 Beginning the WoE process

During the first step of an assessment, whether it is a hazard or risk assessment, the problem and associated questions must be clearly formulated. This includes developing a conceptual model that links the key ecosystem attributes, natural and anthropogenic stressors with important biological receptors and ecosystem services. After this stage, it is then possible to select the appropriate combination of LoE values to answer the problem questions in a systematic, WoE approach. The LoE values can be categorised into three types of information that they provide:

- site characterisation
- exposure characterisation
- effects characterisation.

For example, measuring grain size distributions, temperature, alkalinity and hardness at an aquatic site helps characterise site conditions, which thereby will aid comparisons

between sites. These LoE values should then be combined with exposure characterisation LoE values, such as tissue residues and stressor concentration gradients, which are then closely linked to the resulting biological effect LoE values, such as population indices and toxicity.

The certainty of any WoE conclusion is heavily dependent on five factors:

- the quality of the data
- the quality of the study design
- the expertise of the principal investigator(s)
- the severity of the impairment (ie greater impairment is easier to detect)
- a matching of objectives and data.

Determining the severity of impairment is problematic because it is a relative decision process, determined by what a risk assessor is using as a reference condition and what biological response is being measured. In addition, each LoE may give a different determination of severity because each one measures a different endpoint. Perhaps, most important in the WoE process are Factors 1–3.

Because each LoE provides unique information, and because each has its own associated strengths and limitations, it is important that these traits be considered in the WoE selection and decision process (Burton et al 2005). For example, if the WoE process collects data on metal and chlorinated hydrocarbon concentrations, laboratory toxicity and benthic macroinvertebrate indices at one point in time with split samples, then the strengths of the study are a knowledge of common stressor exposures at one point in time with associated short-term toxicity in constant laboratory exposures and long-term responses of the benthic community. The weaknesses of the approach are a lack of temporal (pulse vs long-term fluctuating) characterisation of stressor exposures and receptor responses; no information on natural stressors, such as habitat, potential artefacts associated with laboratory toxicity exposures; no sublethal effects data; and no documentation of chemical stressors other than metals and chlorinated hydrocarbons. So, this common “triad”-based WoE study had several important attributes, but also some critical weaknesses. This study allows for a WoE decision, but has associated uncertainties that could be reduced with a more comprehensive WoE study design.

1.2 Qualitative to semi-quantitative approaches to WoE

The majority of peer-reviewed studies that suggest they are using WoE simply combine the LoE values in a non-quantitative manner (Burton et al 2002a). These approaches often involve making the decision of effect or no-effect based on the simple majority of the data responses. These approaches are greatly limited by the lack of spatial–temporal information linking exposure and effects.

Many ranking methods have been reported that can be used for WoE-based decisions (eg see review by Swanson and Socha (1997)). Most are based on expert opinions concerning key traits, eg chemical persistence, bioaccumulation or toxicity, and are heavily based on data that are available for chemicals. These approaches may use a hazard index and quotient approach and may even consider chemical mechanisms and interactions (De Rosa et al 1996; Pohl et al 1999; Mumtaz et al 1998); however, they lack site validation. These approaches are sometimes useful in determining highly contaminated sites but less useful for intermediate sites (Bombardier and Blaise 2000), where bioavailability, exposure dynamics and ecological considerations become more important in the LoE profile.

It has become increasingly evident in North America that it is important to include all the appropriate “stakeholders” to be involved in the assessment process. It is even required in the US Environmental Protection Agency guidance (USEPA 2000) and has been re-emphasised in reports by the US National Research Council (National Research Council 2001). Viewpoints on the most important resources, environmental services and biological receptors often vary between stakeholders; therefore, decisions as to whether or not impairment exists needs the input of all, particularly at the early stages of problem and conceptual model formulation. A common method to inform stakeholders is to use ranking methods that involve a consensus-based process (eg Swanson and Socha 1997; Menzie et al 1996, 2000). The Menzie et al approach gives some data more weight than others on a 1 to 5 scale and then placed into a matrix table for evaluation. Results for each measurement endpoint can then be scored in regards to their related assessment endpoint for evidence of harm and then used in the decision-making process. This WoE process also gives greater weight to endpoint concurrence (agreement), which is often perceived as a problem (Chapman 1990, 1996, 2000).

However, the lack of concurrence among the various LoE values does not necessarily mean one is inaccurate; rather, this result may simply reflect the complexity of the exposure vs effect relationship (Hall and Giddings 2000).

There are advantages and disadvantages to the consensus-based ranking approach. Obviously, it allows for consensus among all the stakeholders, which will add support to resulting conclusions and management decisions. The primary disadvantage is the fact that the weightings are subjective, which vary in quality and accuracy depending on the expertise and personalities of the stakeholder participants.

Other semi-quantitative approaches evaluated both within LoE values (eg benthic and fish metrics, combining chemical guidelines, bioassay toxic units) and between LoE values (eg Sediment Quality Triad) can be compared with each other by normalising (eg to percentiles). A more quantitative approach used by Soucek et al (2000) and Cherry et al (2001) combined stream biotic metrics with chemistry and toxicity testing to select the 10 best measures for a model (using statistical methods bivariate and stepwise multiple linear regression analysis), which were scored on a scale of 1 to 10 and then combined to form a hazard rank.

The Sediment Quality Triad has been widely used in the past three decade and was first published by Long and Chapman (1985). Most studies using this method have focused on three LoE values:

- chemical contaminants
- laboratory-based toxicity to surrogate organisms
- indigenous biota community characterisation.

In more recent years, the triad has been expanded to include other LoE values, such as bioaccumulation (Chapman et al 1997; Borgmann et al 2001). Data from each LoE are compared by normalising to reference site conditions on a percentage basis and presented graphically together in the form of a triangle (Chapman et al 1987; Chapman 1990). This presentation allowed the user to compare sites and LoE values; however, because of information loss in the presentation, it has been replaced by more quantitative approaches, eg statistical Spearman rank correlations, principal component analysis, scatterplots (Anderson et al 2001; Read et al 2000; Green et al 1993;

Chapman et al 1996; Chapman 1996, 2000) or by simple matrix tables (Grapentine et al 2002).

The advantage of tabular decision matrices is their ability to clearly present the final LoE and WoE findings, with associated conclusions of stressors and receptor impairment. The matrices can provide information on individual LoE values at a binary classification (eg toxic or not, contaminated or not, altered or not), or use greater levels of discrimination using more powerful statistics. Supporting tables for individual LoE values can provide detailed information.

1.3 Establishing causality in WoE values

There are several recent examples of more advanced WoE values that attempt to better establish causality. Lowell et al (2000) and Culp et al (2000) demonstrated how knowledge gained through time from a variety of sites, with similar stressors (eg pulp and paper mill and municipal wastewater effluents), comprising multiple LoE values, can be used to:

- establish causality
- define acceptable limits
- link environmental components (via weighting LoE component results and summing) in a decision-making framework.

As with any WoE study, essential to the process is high-quality expertise and data from well-designed studies with key LoE values. This allows for the formulation and testing of hypotheses. Causality can be established by following criteria established in the field of epidemiology (Beyers 1998; Fox 1991; Hill 1965; Suter et al 2002) in which some criteria may be given greater importance (ie greater weight) than others.

These approaches establish causality through various mechanisms, such as spatial and temporal correlations of stressor and effect, plausible mechanisms of stressor and effect link, experimental verification of link, strength and specificity of link, biomarker evidence, consistency of link over large geographic area and coherence of link in other regions. USEPA (2000) and Suter et al (2002) provide guidance on establishing causality by 10 evaluations, similar to Lowell et al (2000):

- co-occurrence
- temporality
- biological gradient
- complete exposure pathway
- consistency of association
- experimental confirmation
- plausibility
- specificity
- analogy
- predictive performance.

These determinations are obviously subjective, thus subject to challenge. The determinations are summarised in a matrix table and converted to ranks (eg 1 to 4 or "+" and "-" values). Environmental threshold or effect limits are determined, but these limits are site-specific rather than generic, as they are affected by study design (eg statistical power of test, characterisation accuracy), societal values, and understanding of ecosystem components, dynamics and inter-relationships.

1.4 Quantitative analyses

Quantitative WoE assessments attempting to determine impairment require interpretation of statistical significance, which raises issues of the appropriateness of the

p-value, hypothesis testing, statistical assumptions and the appropriate test (Goodman 1999a, 1999b). Usually, there are multiple statistical methods that can be used where the data meet the required assumptions. Typically, the question is whether or not the system is impaired or not. Most tests are then based on the evidence, and how likely the evidence is, given the possible state of the system using P (evidence-given hypothesis). One computes the probability of observing a more extreme mean than the one observed, ie the p-value, and this is treated as a measure of the strength of the evidence for the (null) hypothesis. This approach is intuitive but may lead to misinterpretation. Users may interpret the calculated probability as the probability that the hypothesis is true, not as the probability of observing the evidence given the hypothesis. Also, the focus is on the null hypothesis and not on how likely the data are, given the alternative, and the sample size vs p-value.

If there are two possible hypotheses, then there are errors that occur when these are incorrectly selected (referred to as the Type I and II error rates). The usual approach is to select a Type I error rate that is small, then select the sample size to make the Type II error rates small for reasonable violations of the null hypothesis. When the null hypothesis is no impact, it is possible to plan a study with a low probability of detecting a problem by selecting a poor alternative, by making the variance in the data large or by not matching the study design to the hypotheses (Peterman 1990).

Some of these difficulties can be overcome using the Bayesian approach, where data are treated as fixed and the parameters as unknowns that have distributions. Because they have distributions, it makes sense to calculate probabilities associated with parameters. This approach leads to a Bayes factor that is the likelihood ratio times the ratio of prior probabilities. Bayes factors are a form of WoE requiring additional evidence of background knowledge to be transformed into inferences and probabilities that a given conclusion is right or wrong (Goodman 1999a, 1999b).

Quantitative statistical approaches allow for combining information to estimate an impairment probability or a measure of impairment. Rules from probability allow for the combination of p-values from individual LoE values (Bailer et al 2002). Multivariate methods are often used as a method for obtaining measures or metrics that may directly convert measurements from multiple LoE values into a single combined LoE. For

example, in principal components analysis (PCA) measurements on many chemicals may sometimes be converted into a single chemical component that may then be used to evaluate impairment. Multivariate methods also provide tools for working with distance-based methods for calculating impairment probabilities, especially useful when the information in LoE values is correlated, such as employed in the BEAST (**BE**nthic **A**ssessment of **S**ediment**T**) (Reynoldson et al 1995). Examples of combining information are estimates of no-observed-effect levels (NOELs), indices of biological status and multivariate measures of effect. The quantitative approach is often based on weighted averages of available information. Two examples of this approach are forming summary measures in PCA or other statistical analyses and forming indicator measures of ecological well-being. Another application where the weighted functions are used is in the estimation of standards or threshold levels. A common approach is to form a weighted average of the NOEL values where the weights are based on the variance (uncertainty) of the estimated value. Another example is the calculation of threshold levels to protect ecosystems based on species sensitivity distributions. Toxicity thresholds on multiple species are combined, usually using an ordinary mean or mean of the log transformed data. A lower percentile is then calculated based on an assumed distribution (Posthuma et al 2001).

A useful visual and statistically based WoE approach combines Geographic Information Systems (GIS's) with WoE (Kapo and Burton 2006). This technique employs a data-driven, probabilistic approach originally developed for use in minerals exploration. This GIS-based WoE method is superior to qualitative methods commonly used in typical WoE assessments as it provides a less-biased weighting approach. Weights are assigned to ecosystem stressors based on quantitative spatial correlation, rather than expert opinion, and represent actual field observation responses. The analysis takes into account the important role of stressor mixtures in ecosystem impairment, including the ability to identify cases where chemicals present at low concentrations may have a detrimental effect in the presence of one or more other environmental conditions. A case study (Kapo and Burton 2006) yielded baseline results consistent with other assessments (de Zwart et al 2006), as well as providing additional results quantitatively ranking watershed stressor contributions and probability of impairment at designated spatial resolutions (Figures 1–3). Both existing data and hypothetical data (eg post-remedial sediment contaminant concentrations, changes in point-source outputs) can be

used in the creation of probabilistic impairment maps, and models can include multiple species. This approach can be used to integrate the results from the various LoE values collected at the site to allow for improved decision making on risk, recovery and remediation effectiveness.

Fig. 1a. IBI "Impaired" Sites (<40)

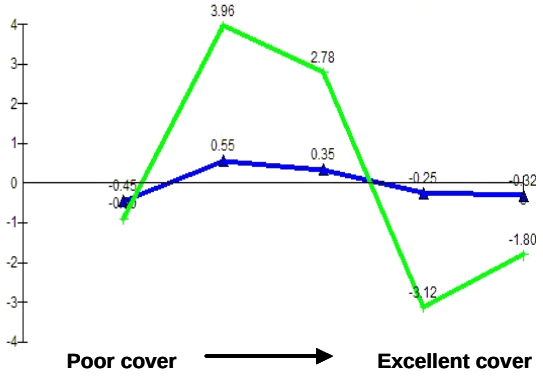

Fig 1b. IBI "Good" Quality Sites (>44)

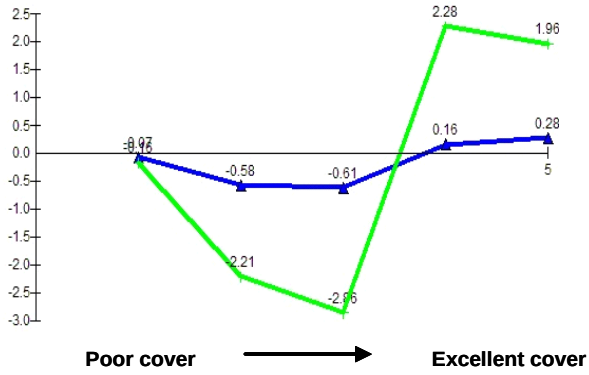

Fig 2a. IBI "Impaired" Sites (<40)

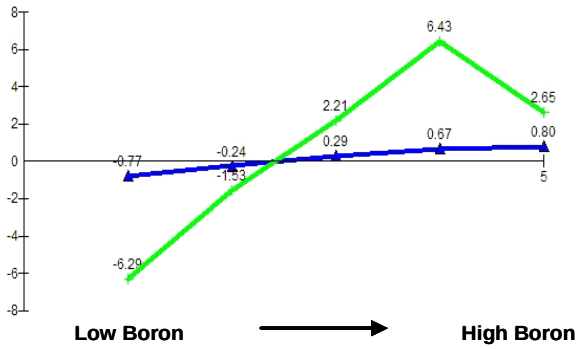

Fig 2b. IBI "Good" Quality Sites (>44)

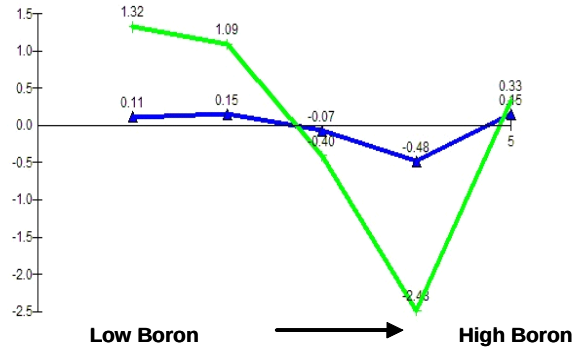

Figure 1: Example of the relationship of habitat cover to fish impairment in the Great Miami River watershed

Response of site occurrence for (a) "impaired" Index of Biotic Integrity (IBI) sites (IBI < 40) and (b) "good" quality IBI sites (IBI > 44) to a five-level gradient of cover habitat conditions, from existing Qualitative Habitat Evaluation Index (QHEI) data for cover.

Figure 2: Example of the relationship of boron concentrations (possible municipal wastewater surrogate) in water to fish impairment

Response of site occurrence for (a) "impaired" IBI sites (< 40) and (b) "good" quality IBI (> 44) sites to a five-level gradient of boron concentrations (ug/L).

In Figures 1 and 2, blue lines indicate the weight of positive spatial association at each level, while green lines represent the overall confidence level for the weight. A confidence value of 1.96 corresponds to 95% confidence (Kapo and Burton 2006).

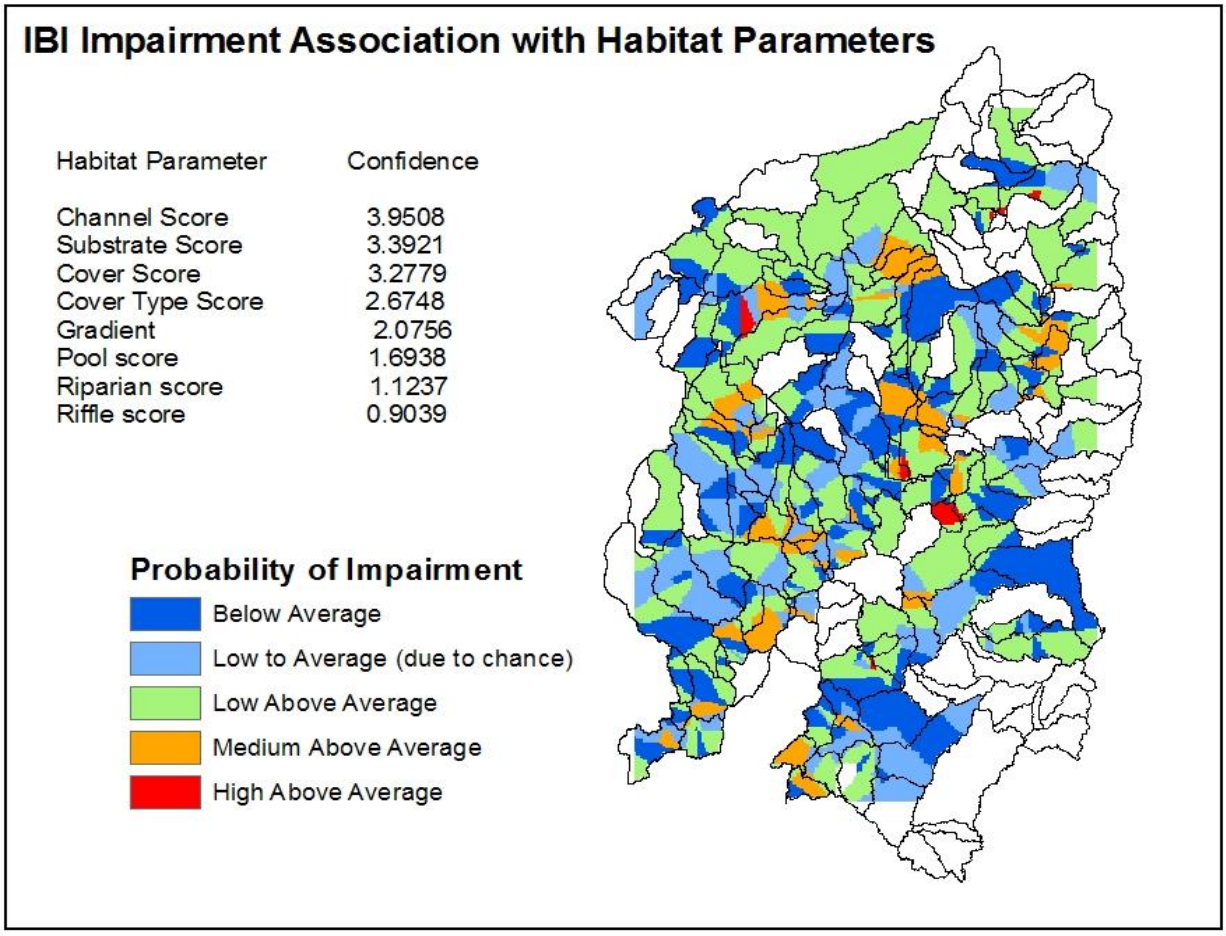

Figure 3: Example of weighted logistic regression probability map for impairment of fish communities using the IBI, due to habitat characteristics only (Great Miami River watershed, divided into sub-basins for viewing purposes using ArcHydro) (Kapo and Burton 2006)

2. THE WOE FRAMEWORK FOR DETECTING OR PREDICTING ECOSYSTEM IMPAIRMENT OF MULTIPLE STRESSORS

The recommended WoE framework follows Burton et al (2002b), as illustrated in Figure 4. This is a process that allows a wide variety of study designs and methods to be utilised, and is most applicable to site-specific and regional assessments of impairment or risk.

Each of the framework components (boxes) is considered a certainty element that should be addressed in a transparent process. At the beginning of the assessment (eg problem formulation stage), the first three certainty elements are carried out. These steps then allow for development of the conceptual model.

2.1 Certainty Element 1: Critical biota

The most important biota (eg an organism, population or community; threatened or endangered species; keystone species; commercially important species; species of stakeholder concern) must be identified for the study system and at appropriate reference sites. Their selection will depend on stakeholder and ecological concerns (eg sensitivity, sentinel or threatened status, abundance, value), ability to measure and the eco-geographic region.

2.2 Certainty Element 2: Ecosystem quality

The term quality is very subjective and relative. Stakeholders must agree on what defines high quality for their area of concern. This process is closely linked to a subsequent certainty element of reference site selection. The basis of all decisions regarding impairment is in comparison to a reference, benchmark and/or control value.

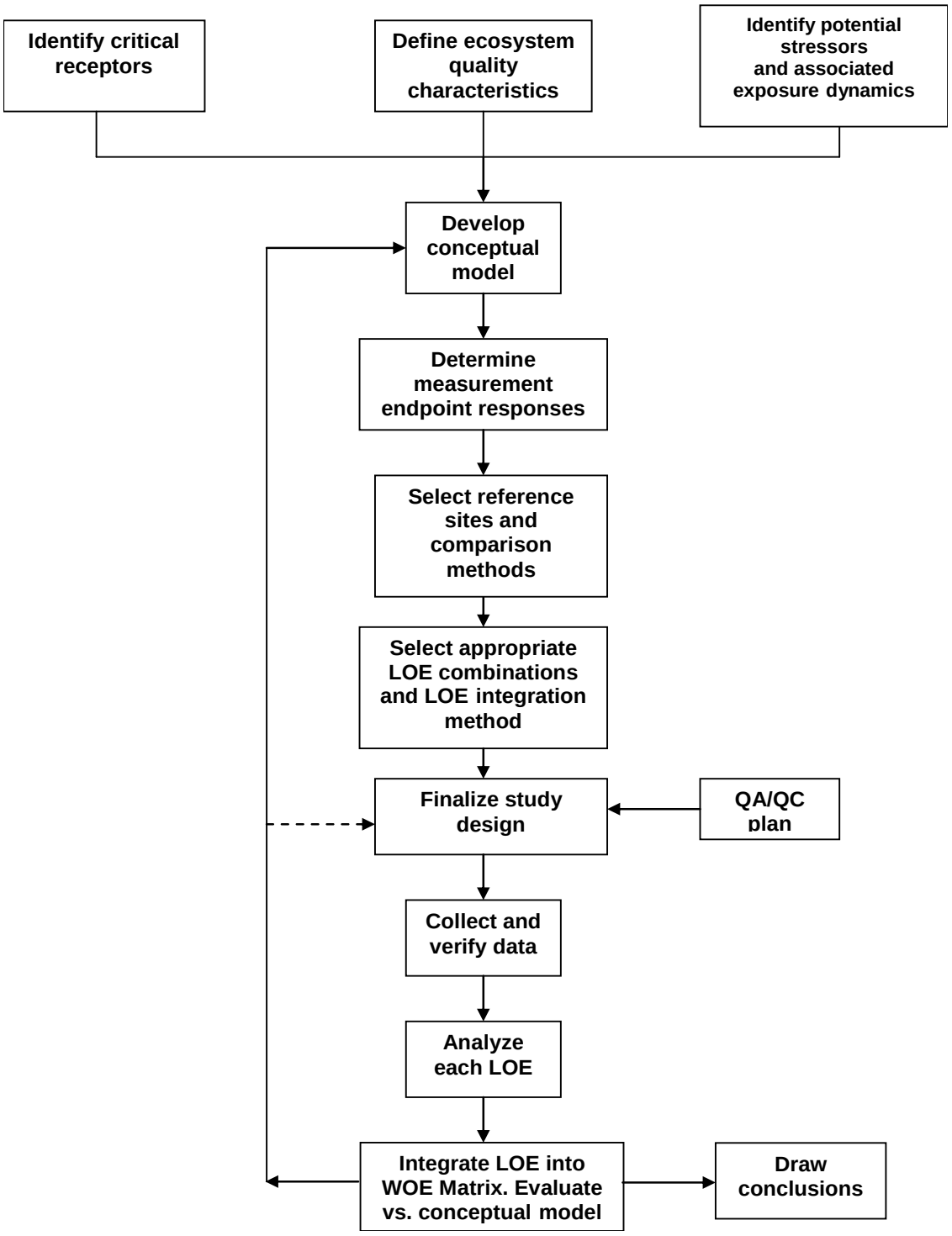

Figure 4: Critical elements (CEs) of the weight-of-evidence (WoE) framework (Burton et al 2002b)

2.3 Certainty Element 3: Potential stressors and their distribution

Ideally, during the initial risk assessment problem formulation, the potentially dominant stressors (chemical, physical and biological) that affect the critical biota need to be identified, with a general understanding of their spatial and temporal distributions. This will be critical to designing the sampling plan. A pilot (pre-screening) study would help define this distribution and associated variance.

2.4 Certainty Element 4: Developing a conceptual model

The conceptual model should link critical receptors, ecosystem quality characteristics, and all possible natural and anthropogenic stressors in addition to considering the associated exposure dynamics. As part of the model development, the links between model components and measurement endpoint responses (direct and indirect responses with associated spatial/temporal dynamics) need to be defined and will dictate which LoE is selected.

2.5 Certainty Element 5: Select measurement endpoint responses

Critical receptors (eg target species such as mayfly nymphs as food for trout) within the conceptual model must be measured either directly, or indirectly using surrogate species. Surrogates should be selected based on relevant life histories, ecosystem characteristics and other concerns (eg productive fishery, pathogens). An important aspect of this certainty element is the need for adequate representation of critical components of the conceptual model that can distinguish among alternative hypotheses. For example, endpoints from multiple exposure characterisation LoE values should be used to account for chemical and physical factors that strongly influence mayfly population dynamics (eg resident populations, habitat, water quality and toxicity measures) to improve causality determinations. Consider strengths and limitations of site characterisation, exposure characterisation and effects characterisation LoE values and then select so that each LoE contributes to the WoE decision-making process that links stressor exposures with adverse effects (see above discussion). Also ensure that all LoE values meet quality assurance/quality control (QA/QC) criteria including the use of validated methods. Without quality data, uncertainty increases markedly.

2.6 Certainty Element 6: Select appropriate reference comparison

Perhaps the most important, and difficult, aspect of any impairment/risk assessment is selection of the appropriate reference site. Because no two sites are the same, it is more appropriate to define a reference condition, defining a range of conditions that is appropriate to represent “least impaired” conditions relevant to your site and associated land uses. The most appropriate reference or references can vary widely between studies and depend on a host of factors, such as study objectives, receptor life histories, land use and watershed impairment. However, given the natural variability of reference site ecosystems (eg Hawkins et al 2000), their rarity in many human-dominated watersheds and study resource limitations, it is often difficult to characterise an appropriate reference condition.

(Paragraph break here)

The number of reference sites chosen will depend upon available resources, geographical scale and habitat heterogeneity. Individual site groups should aim for a minimum of five and preferably at least 10 sites (Wright 1995). The commonly used upstream reference may not be appropriate for defining impairment as it is not truly independent from downstream sites and may differ significantly in its natural characteristics (eg substrate, flow and temperature). In streams full of effluent or in those where upstream reference conditions are ephemeral/intermittent, downstream reference sites may be appropriate if the biotic communities and water quality characteristics represent unimpacted conditions. Landis and McLaughlin (2000) suggest use of a reference space or condition, utilising multiple LoE values, to define an assessment space. Hydrologic and habitat characteristics must be strongly considered in the selection of appropriate reference sites.

2.7 Certainty Element 7: LoE selection and their integration

As discussed above, it is important to recognise the advantages and limitations of each LoE and select an appropriate and complementary combination of LoE values. Included in the analysis should be the causality criteria used for each LoE during output verification (see Figure 4). Almost invariably, there should be a minimum of three LoE

values that characterise stressors and link to effects. Effects-based LoE values are the most important, because they are direct measures of what is being protected.

The selection of the appropriate statistical methods to integrate the data derived from the various LoE values must be considered in light of the sampling design. For example, appropriate statistical quantification methods for different impairment categories include:

- application of regression analyses across an impairment gradient
- application of analysis of variance (ANOVA) and before/after analyses to quantify response following a discrete impact event
- application of ANOVA and other multivariate methods using multiple reference sites
- application of regression, ANOVA and/or other multivariate methods where gradients and reference sites exist.

2.8 Certainty Element 8: Study design and QA/QC plan

The study design should link the conceptual model components with the data that are collected. For example, if sampling along a gradient of stress and responses is used, then a regression model is needed. For a paired analysis, emphasis needs to be placed on the selection and matching of the pairs, the number of pairs and potential confounding factors that might affect the pairs. Essential study design elements will include, but are not restricted to:

- selection of optimal measurement endpoints
- laboratory and field quality control procedures to ensure observations with the lowest possible variability
- selection of appropriate model relating stress and response
- characterisation of dominant natural and anthropogenic stressors and gradients
- selection of appropriate reference sites and background stressor levels
- selection of adequate sampling sites and numbers of samples to ensure adequate statistical power to detect pre-specified biologically significant changes in responses and spatial/temporal variance
- identification of key exposure pathways, eg low/high flow, ground water upwelling and surface water down welling (GW/GS), sediment (surface and deep), food

- characterisation of effects using biologically based methods, eg resident biota and toxicity
- selection of appropriate statistical methods for LoE analyses and integration into a WoE matrix.

2.9 Certainty Element 9: Data collection and verification

In order to better link stressor exposures with adverse effects, the various LoE values should be collected in a synoptic, contemporaneous fashion to increase the certainty of exposure–effect linkages when they are integrated. Beyond routine QA/QC data, data validation requires expert review of data vs sampling protocols to identify any questionable findings. Questionable data should be labelled with appropriate qualifiers. By addressing the following 13 data questions, one can minimise the uncertainties associated with the project conclusion:

- Do chemical data represent bioavailable fractions?
- Are spatial/temporal stressor exposures characterised (eg hydrodynamic influences)?
- Are major stressors (natural and anthropogenic) identified and considered?
- Are tissue residues of contaminants measured in species with a relevant home range?
- Are field data subject to confounding influences (eg do indigenous biota responses consider natural/habitat stressors?)?
- Are there suspicious outliers and should they be included/excluded?
- Can observed effects be explained by known stressors?
- Are stressor interactions (eg antagonism, synergism) considered?
- Are laboratory and field sampling artefacts known?
- Are reference sites appropriate?
- Are taxonomic identifications sufficient?
- Are field observations appropriate?
- Were data collected during optimal times?

2.10 Certainty Element 10: Analysing LoE

Data within each LoE should be analysed by the most appropriate and optimal quantitative methods so that statistical significance and associated variance is known. It is preferable not to integrate data such that individual LoE information is lost; however, it may be useful on occasion. An example would be consolidation of benthic invertebrate data into various indices (eg diversity, taxa richness) and then combining the indices into an overall site metric (eg Index of Community Invertebrates, Index of Biotic Integrity, Index of Well-Being). Aspects of the analyses that must be addressed and documented include, but are not restricted to, QA/QC; stressor magnitude, frequency, duration and interactions; and exposure–biological effect relationships. Estimation and evaluation of uncertainty of these aspects is a critical component of analysis, such as is done in probabilistic risk assessments. The uncertainty assessment might involve estimation of variance in response, effects of choice of model and associated assumptions, sensitivity analysis of model components and perhaps a qualitative evaluation of all aspects of the study design. Causality criteria (see 1.3 above discussion) used to link stressors and effects (for each LoE) should be clearly stated, together with how the links were established or refuted (Burton et al 2002b). By connecting the method of data integration to the study objectives and available information, it is often possible to select an appropriate method of integrating the information in a single LoE (Smith et al 2002).

2.11 Certainty Element 11: Integrating LoEs into a WoE

Multiple LoE values may be integrated if done in a contemporaneous fashion. This can be useful when integrating multiple LoE values. For example, the BEAST approach uses multivariate statistics to combine measures of sediment physico-chemistry, benthic communities and sediment toxicity to identify reference vs impaired conditions (Reynoldson et al 1995; Reynoldson and Wright 2000; Grapentine et al 2002). This approach requires a large database, effectively defining a reference condition. Statistical differences between biotic responses and habitat and chemistry conditions can thereby define impairment. Conversely, when multiple LoE values are collapsed into a single number, data are lost and may oversimplify the evidence. Examples of WoE integration of various LoE values are presented by Bailer et al (2002), Smith et al (2002),

Grapentine et al (2002) and Forbes and Calow (2002). These authors used a variety of acceptable approaches that meet the WoE framework guidelines described above.

The framework described is useful for establishing impairment and risk and provides for a mechanism for establishing causality (discussed above). Expert qualitative judgements on causality can be summarised in a tabular decision matrix, for example converting to ranks (eg 1 to 4 or "+" and "-" values), as shown by Chapman (1990, 1996), Grapentine et al (2002) and USEPA (2000). This approach can be useful to both experts and stakeholders, but requires that the WoE critical elements be conducted in a thorough, high-quality manner and the conclusions from the individual LoE values be quantitatively based, with an adequate level of power. This then allows an interdisciplinary team to combine the LoE values into a WoE matrix table for the decision-making process (eg Grapentine et al 2002).

Preferably, the WoE analysis considers attribute weightings and causality evidence in a consensus-based approach that has a strong statistical basis and involves stakeholders and experts. Once the quantitative analyses have been carried out, the tabular decision matrix table should be revisited. The importance of each LoE should be considered by ranking uncertainty, ecological relevance, societal importance and the size of the ecological change (Smith 2002). Another critical component associated with evaluating evidence is the role of error rates in the decision process. As a result, the hypothesis-based assessment process is frequently not protective of the environment (Peterman 1990). The problem is not with the hypothesis test but rather with decisions that are made about the hypotheses and error rates. Alternatives to the basic approach are to swap the null and alternative hypothesis or to decide to balance the Type I and Type II error rates. A sensitivity analysis within appropriate LoE values is also important to ensure that the critical aspects of any model can be identified and the uncertainty in those parameters reduced. Decisions about these analytical options need to be made during the problem formulation stage of the assessment (Burton et al 2002b).

3. INTEGRATION OF WOE CONCEPT INTO THE RISK EVALUATION PROCESS OF SINGLE STRESSORS

The WoE concept is embedded in several risk assessment frameworks dealing with the evaluation of potential risks posed by single substances. Within the EU REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) legislation, the so-called WoE approach is a component of the decision-making procedure on substance properties and thus an important part of the chemical safety assessment (CSA). In the legal text, the use of the WoE approach is provided for in Annex XI as an option to meet the information requirements of Annexes VII to X. Here it is used to give a rationale to show that the compiled data adequately describe the REACH endpoint of concern and that further information on that particular endpoint may not be necessary. As such, WoE is here closely linked to integrated testing/information strategies (ECHA 2010).

However, the WoE concept could be used in any aspect of the risk assessment process involving determinations of exposure (bioavailability and physico-chemical conditions, biogeochemical or metallo-regions) and effects (laboratory and field data on species, populations and communities), which are then linked in the risk characterisation process (see the MERAG Fact Sheet 1 – Figure 1: Generic tiered approach for performing the risk characterisation for metals/metal compounds). The WoE could involve single aspects of the assessment process, such as deterministic, probabilistic and field validation studies. Finally, all of these components can be integrated into a full WoE decision-making process.

The European risk assessment approach has as its foundation the use of predicted no-effect concentrations (PNEC), which have increased certainty based on the tiered evaluation that ensures data quality and relevance, a focus on bioavailable fractions of metals and transparent data analyses with associated statistical confidence. This approach is useful for continental and regional management applications. However, as the geographic area of focus narrows (such as from continental to regional/ecoregion to sub-ecoregions to site-specific), the uncertainty of the PNEC approach increases as the relevant database (including local organisms and physicochemical conditions) decreases. This increasing uncertainty can be countered with additional LoE values that incorporate a field validation component, thereby allowing for the refinement of predictive

models (eg acid-volatile sulfide (AVS), total organic carbon (TOC) bioavailability normalisation factors, indigenous species toxicity thresholds, population/community indices). For additional information on utilising WoE in the field assessment process, see the discussion in the preceding sections.

3.1 Using WoE in the PNEC process

3.1.1 Considering all LoE values

In Europe, it is common to derive a PNEC when establishing safe levels (chemical standards) of chemical exposure in water, soil and sediment. This is also an integral component of the MERAG process and is explained in detail in Fact Sheet 8 (classification of metals, alloys and complex materials) and Fact Sheet 3 (effects assessment) for use in the effects assessment. The process for this derivation varies and can utilise various LoE values, such as chronic ecotoxicity data on species, field data and theoretical models (such as equilibrium partitioning, biotic ligand model).

As discussed in the WoE framework process above, the quality of the data, their realism (lack of uncertainty) and the degree to which the ecosystem and relevant receptor responses have been characterised are all important issues to consider in the PNEC derivation process. For example, there is substantial uncertainty associated with PNEC values only derived from laboratory toxicity data on a few surrogate species, due to extrapolating to field conditions where bioavailability and exposure dynamics differ, and where there are a myriad of indigenous species with both direct and indirect interactions. There is also substantial uncertainty using only field-derived empirical guidelines (such as probable effect levels (MacDonald et al 2000)). Here, data from a multitude of sites, where there are multiple contaminants, are used to predict single chemical threshold effects at other sites, leading to unacceptably large uncertainties (over 30%) and requiring the use of additional LoE values (Adams et al 2005). Therefore, to reduce this uncertainty, additional LoE values should be utilised in a sound, relevant and statistically powerful manner (Wenning et al 2005).

The combination of only two LoE values, such as laboratory toxicity data with field-derived effects data, however, greatly reduces uncertainty and adds credibility to the PNEC derivation process. This credibility is strengthened with the addition of other

useful LoE values such as theoretically sound and proven approaches, for example organic carbon and AVS normalisations of sediment chemistry data for non-polar organics and divalent metal cations, respectively, or the use of benthic species, population and community chronic responses to spiked sediments over long periods of time (Burton et al 2005, 2006; DiToro et al 1985, 2005). Sediment spiking studies are useful LoE values if conducted properly. The two primary problems that occur in spiking studies are lack of adequate equilibration time (ie partitioning of chemical to sediments), which can greatly increase organism exposure through the water route, and, a related problem, the inability to separate water column exposures and effects from those occurring in the sediments. Several studies have shown that several weeks of equilibration may be required for sediments spiked with metals (eg Simpson et al 2005). The advantages of the spiking study are that multiple concentrations can be used to derive more accurate threshold levels with associated known variance and they can be utilised in field settings, thereby allowing for exposures to indigenous populations under natural conditions. Other useful LoE values include tissue residue data for contaminants that have been linked to adverse effects, as long as there is a clear relationship between exposure and effects. These data can be problematic due to multiple exposure pathways (eg overlying water, sediment, food) not being clearly separated. In addition, there are interpretation difficulties when the metals are essential elements (eg copper and zinc). Accumulation of metal may not equate to adverse effects.

3.1.2 Data quality, quantity and relevance

Data compilation and selection issues are discussed in detail in the MERAG Fact Sheet 3. The quality of the data utilised in the PNEC derivation requires a screening process to which all parties (stakeholders) agree. Preferably, this process and the criteria for data acceptance or rejection would occur prior to the assessment process, during the problem formulation stage. This process should include several QA/QC checks prior to an evaluation for relevance (see below). For laboratory toxicity tests that are standardised, there are accepted performance criteria for testing conditions, such as test organism type, organism age and health, replicate numbers, range of physico-chemical conditions (such as pH, dissolved oxygen, salinity, hardness and alkalinity, temperature) and adequate control performance. For bioaccumulation – tissue data derived from laboratory exposures – there must be additional QA/QC checks for the analytical components of the study, including replication and associated variance, spike recovery

and minimum detection limits. Field data, such as empirical data linking concentrations in sediments vs indigenous responses, or tissue residue data have some similar and differing QA/QC checks. For example, the analytical QA/QC requirements are similar, but the need for additional supporting data that document exposures and bioavailability are needed (such as AVS, TOC, dissolved organic carbon (DOC), total suspended solids (TSS) and other contaminants). Also questions of adequate replication, sample collection and manipulation (such as reducing the loss of AVS), and sample transport and storage time and conditions, can greatly influence the quality of the data (Environment Canada 1995; USEPA 2000).

Data quantity can also be an extremely important issue that is not always given adequate consideration in the WoE process. Uncertainty tends to increase with a decreasing “n” due to basic statistical assumptions. The variance (eg confidence intervals) increases and makes derivation of threshold levels (eg PNEC values) problematic and inaccurate. However, more data do not necessarily equate to better conclusions, as the data should be of high quality, relevant and show a clear linkage between exposure and effects. One of the primary limitations of biological effects (toxicity) data is the lack of spatial and temporal characterisation of exposure under realistic conditions (see above discussions on LoE advantages and limitations). This issue is seldom discussed because adequate exposure data are seldom available. So there is a data quantity issue that cannot be separated from data relevance issues.

Determining data relevance can be difficult and very subjective, and also requires stakeholders to discuss acceptance criteria at the beginning of the process so that hopefully consensus can be achieved prior to data analyses. Species sensitivity distributions (SSD) have become popular in recent years and have several advantages for determining threshold effect levels (Posthuma et al 2001). This allows for a probabilistic approach to determine a concentration that will protect most species in the ecosystem, but this assumption explicitly depends on other assumptions. The SSD predicted threshold effect determination becomes more reliable with increasing numbers of species and toxicity data. However, the data included in the process may: 1) be non-indigenous to the ecosystems being managed, 2) have questionable QA/QC and 3) have been derived in laboratory exposures that differ widely between the reported species, with varying statistical power and under conditions that are irrelevant to the ecosystems

in question. Finally, the question of how well these laboratory data on single species, under constant exposures in “clean” systems, relate to natural exposures (fluctuating concentrations, differing bioavailability due to “dirty” components) and to indirect effects, populations and communities is unknown. These significant unknowns are limitations of this useful LoE, thus requiring additional LoE values that are more field based.

3.1.3 Derivation of a PNEC

Utilisation of the PNEC in a WoE process is also discussed in the MERAG Fact Sheet 3. It is apparent from the many issues discussed above that address improving certainty in the determination of threshold effects, ie PNEC values (thus protecting the ecosystem from adverse harm), will require multiple LoE values, and should be integrated in a WoE process. The key LoE values of greatest weight (significance) would be those that best link exposure with effects under realistic conditions, with the fewest assumptions and uncertainties. For example, field-based data that describe exposure to the chemical of concern – over a wide range of conditions, for long time periods – that is linked to effects in multiple indigenous species and receptors of concern could have significantly more weight than constant laboratory-based exposures to only a few surrogate species. The challenge in the field exposure data, however, is to have an adequate experimental design with enough statistical power to derive a PNEC that has acceptable variance. Often, the study may produce unbounded no-observed-effect concentration (NOEC) and lowest-observed-effect concentration (LOEC) values. So the field data are then best used to provide a crude validation of theoretical predictions or bioavailability and/or laboratory-derived thresholds. It should, however, be noted that interpretation of field data can be difficult. A problem associated with field-based data is that they are subject to many confounding factors, for instance the presence of other contaminants, difference in bioavailability from changing physical/chemical toxicity modifying factors (eg water hardness), seasonal variability, etc. Therefore, it might be difficult to attribute observations (effect) to only one stressor or contaminant. Thus, multiple LoE values become essential, combining the field, laboratory and modelling components to better derive PNEC values that are bioavailability- based, and using bioavailability normalisation factors and species that are relevant to the ecosystems of concern (see Example 1).

Example 1: Use of the weight-of-evidence (WoE) approach in setting a PNEC freshwater sediment for copper compounds

In April 2008, the copper industry completed a voluntary risk assessment on copper, including the derivation of a predicted no-effect concentration (PNEC) for copper in freshwater sediments. The risk assessment was agreed by the European Commission and EU Member States. The European Commission's Technical Committee for New and Existing Substances and the European Commission's Scientific Committee on Health and Environmental Risks (SCHER) also performed an evaluation and further endorsed the conclusions on the environmental and human-health risk characterisation. For full details of the risk assessment and the PNEC sediment derivation, the reader is referred to the documents posted at the official website of the European Chemicals Agency (ECHA) (http://echa.europa.eu/chem_data/transit_measures/vrar_en.asp).

Compared to organic substances, copper is a relatively data-rich substance with regard to the availability of freshwater-sediment toxicity data. Nevertheless, the use of a WoE approach was deemed highly relevant in order to make a decision on the appropriate assessment factors (reflecting remaining uncertainty) to be used to derive a final PNEC. Different lines of evidence (LoEs) were used in this regard. Starting point for the initial PNEC derivation was the availability of a data set of benthic toxicity data. Supporting evidence was compiled using different sources and tiers of information: (1) use of pelagic ecotoxicity data in combination with Kd values derived through different approaches, (2) acid-volatile sulfide (AVS) and total organic carbon (TOC) normalisation procedures, (3) use of soil ecotoxicity data and soil bioavailability models and (4) mesocosm/field ecotoxicity data.

LoE 1: Benthic toxicity data

Available single-species sediment exposure tests resulted in 106 individual high-quality chronic no-observed-effect concentration (NOEC) values for six different sediment dwelling organisms, ie the amphipods *Hyalella azteca* (25 individual NOEC values) and *Gammarus pulex* (6 individual NOEC values), the oligochaetes *Tubifex tubifex* (39 individual NOEC values) and *Lumbriculus variegatus* (3 individual NOEC values), the midge fly insect *Chironomus riparius* (27 individual NOEC values) and the mayfly insect *Hexagenia* (6 NOEC values). The selected NOEC values of the non-normalised data set ranged between 18.3 mg/kg dry weight and > 3,158 mg/kg (min-max value) and exhibited also large intra-species variability. This large variability observed in the reported effect levels was mainly attributed to sediment characteristics such as the amount of organic carbon (OC) present and the presence of sulfides. Since both AVS and OC seem to mitigate copper toxicity, the use of the whole effect database without any form of correction for bioavailability would subsequently lead to erroneous calculations of the PNEC value for

freshwater sediments.

The derivation of the freshwater $HC5\text{-}50_{sediment\ (benthic\ SSD)}$ for copper has thus been based on the OC normalised data set, using only low AVS sediments, and includes six species-specific data points (see above) representing 62 NOEC values. The HC5-50 (range of effect concentration affecting 5% to 50% of organisms) was estimated using the statistical extrapolation methodology. Both best fit distribution and the lognormal distribution were derived.

LoE 2: Use of the equilibrium partitioning (EqP) method

Generally metal toxicity in sediments is mainly exerted via pore water exposure (eg simultaneously extracted metals–acid-volatile sulfide (SEM-AVS) concept). As substantiated by numerous publications, the use of the equilibrium approach can be considered for metals. The application of the EqP approach to the copper data is performed in two steps: (a) using the median dissociation constant (Kd) values obtained from monitoring data and (b) using the Kd values, calculated by the Windermere Humic Aqueous Model (WHAM).

Using the EqP approach, $HC5\text{-}50_{sediment\ (EP)}$ values were derived for seven aquatic EU scenarios, representative for the different physico-chemical characteristics encountered in EU surface waters. The scenario-specific $HC5\text{-}50_{sediment\ (EP)}$ values were calculated from the scenario-specific aquatic HC5-50 values (using 139 NOEC values from 27 species, including seven benthic species) and the application of the following Kd values: the EU median $Kd_{suspended\ solids}$, the EU median $Kd_{sediment}$ and scenario-specific Kd values calculated from WHAM VI Kd ($_{WHAM).}$ These approaches resulted in the $HC5\text{-}50_{sediment\ (EP\ SS),}$ $HC5\text{-}50_{sediment\ (EP\ Sed)}$ and $HC5\text{-}50_{sediment\ (EP\ WHAM).}$ Considering the relevance of OC binding, all values were normalised for their OC content. For each approach, the lowest HC5-50 values (1,833 to 3,808 mg Cu/kg OC) were selected as the reasonable worst-case $HC5\text{-}50_{sediment\ (EP)}$ values.

LoE 3: Mesocosms and field data

In a third tier, threshold values obtained from mesocosms and field data are evaluated and compared to the Tier 1 and Tier 2 derived $HC_{5\text{-}50}$ values. Sediment threshold values and benthic NOEC values are available from the four mesocosm studies and one field cohort study. The studies cover ecotoxicological relevant endpoints for a wide range of taxonomic groups important for the benthic structure as well as the benthic functions (eg microbial degradation). The lowest mesocosm OC normalised NOEC (4,285 mg Cu/kg OC) is a factor 2.1 (best fit) to 2.5 (lognormal) above the $HC5\text{-}50_{sediment\ (benthic\ SSD)}$. An OC-based $HC5\text{-}50_{sediment\ (mesocosm\ SSD)}$ (5^{th} and 95^{th} confidence limits) was calculated as 3,007 mg Cu/kg OC (2,204 to 3,743 mg Cu/kg OC).

LoE 4: Comparison with terrestrial data

Considering freshwater sediments as “wet soils” also allowed for a comparison between the HC5-50 values, derived from sediment NOEC values with OC normalisation, and the HC5-50 values derived from soil NOEC data (251 NOEC values, covering 19 species of plants/invertebrates and nine microbial endpoints) and soil bioavailability models (pH, OC and cation exchange capacity (CEC) normalisations).

Integration of all LoE into the WoE process

The HC5-50 values of all LoE are summarised in Table 1.

Line of evidence	Approach	HC5-50 (organic carbon normalised (mg Cu/kg OC)	Weight
LoE 1: Benthic toxicity data			
	SSD best fit and low AVS	2,021	+++
	SSD lognormal and low AVS	1,741	++
LoE 2: Equilibrium partitioning			
	EP-WHAM RWC	1,833	+
	EP-KdSS RWC	2,359	+
	EP-KdSED RWC	3,808	+
LoE 3: Mesocosms/field data			
	Mesocosm/field SSD lognormal	3,007	++++
LoE 4: Terrestrial data			
	Application soil NOEC values and soil bioavailability models	1,723–3,817	±

Table 1: Overview of all HC5-50 values obtained by the different lines of evidence used

The HC5-50 derived from the benthic toxicity data results in 1,741 mg Cu/kg OC using the lognormal distribution. This value was, subsequently, compared with the other obtained HC5-50 values in a WoE approach to assess the remaining uncertainty. Comparison between the OC normalised HC5-50 values obtained with the EqP approach shows that the $HC5\text{-}50_{sediment\ (EP\ WHAM)}$ values are somewhat lower than $HC5\text{-}50_{sediment\ (EP\ Sed)}$ values. The difference is likely related to the fact that in the WHAM approach only binding to OC is considered, while natural sediments contain additional binding sites (copper binding minerals and AVS). This additional binding contributes to the Kd values reported in the literature; variations in mineral binding, sulfide binding and OC sequestration help explain the high variability of those values in the literature.

The lowest mesocosm OC normalised NOEC (4,285 mg Cu/kg OC) is a factor 2.1 (best fit) to 2.5 (lognormal) above the $HC5\text{-}50_{sediment\ (benthic\ SSD)}$. An OC-based $HC5\text{-}50_{sediment\ (mesocosm\ SSD)}$ (5^{th} and 95^{th} confidence limits) was calculated as 3,007 (2,204–3,743) mg Cu/kg OC. The mesocosm HC5-50 is a factor 1.5 to 1.7 above the derived $HC5\text{-}50_{sediment\ (benthic\ SSD)}$ (not significant at 0.05 level but significant differences at the 0.01 level). The mesocosm data therefore clearly demonstrate that the $HC5\text{-}50_{sediment}$ values, derived through EqP and single-species sediment toxicity testing, are protective for a wide range of benthic organisms, tested in a variety of conditions. The mesocosm validations include multi-exposure routes and multi-species interactions and account for benthic structures as well as functions (including sediment decomposition).

Finally, the comparison for a range of representative sediment scenarios shows that the HC5-50 values estimated from, respectively, sediment and soil data are highly correlated and that the HC5-50 values derived from the sediment NOEC values/OC normalisation were on average between a factor 0.7 to 0.8 below HC5-50 values derived from soil NOECs/bioavailability models (no significant differences could be detected between the OC-based HC5-50 values derived). This comparison therefore adds further evidence regarding the protective nature of the $HC5\text{-}50_{sediment\ (benthic\ SSD)}$ values, derived from benthic ecotoxicity tests.

In order to evaluate the remaining uncertainty, and the need for an additional assessment factor (AF) on the HC5-50, the following elements were also considered in the WoE approach: data quality, taxonomic groups covered, treatment of multiple data sets, statistical uncertainty around the 5^{th} percentile estimate, evaluation of NOEC values below the HC5-50 estimate, comparison with natural background levels and essentiality levels.

The uncertainty analysis revealed that there was no need for applying an additional assessment factor on the HC5 value of 1,742 mg/kg OC, yielding for a sediment of 5% OC a PNEC of 87 mg/kg dry weight.

4. CONCLUSIONS

The WoE framework developed by Burton et al (2002b) was for the assessments of sediment contamination and their effect(s) on ecosystem impairment. However, the framework is applicable to other ecosystem types and assessment processes, such as MERAG. Paramount in the application and usefulness of this WoE framework into the decision-making process is:

- transparency and stakeholder involvement
- strong scientific validity
- ecological relevance with clear linkage of exposure and effects
- known statistical variance in key assessment components.

These issues are particularly important during the initial problem formulation and final risk characterisation stages of the assessment.

The methods employed in the final integration of the various LoE values will vary widely, based on the study design, types of data and availability of quality data. The quantification scenarios for analysing the LoE will depend on the impairment, characteristics of the system and resource availability for the assessment. Typical study design/statistical quantification scenarios are:

- gradient (regression methods)
- paired reference/test (before/after control impact and ANOVA methods)
- multiple reference (ANOVA and multivariate methods)
- gradient with reference (regression, ANOVA and multivariate methods).

Ideally, LoE values will be integrated in a contemporaneous study, to better link exposures with effects, but this is not possible in regional or continental, retrospective studies. More typically, the integration of LoE values will be using a WoE matrix table, which is made up of the individual LoE values that have already been statistically evaluated. Even when the various LoE evaluations are inconsistent with each other, reasonable WoE conclusions are still possible (Grapentine et al 2002). Nevertheless, weighting or integrating multiple LoE values into a conclusion does not remove uncertainty. It can provide a sound, transparent process for reducing uncertainty by integrating the best available scientific information available at the time.

5. REFERENCES

Adams, W J, Green, A S, Ahlf, W et al (2005). Using sediment assessment tools and a weight-of-evidence approach. In RJ Wenning, GE Batley, CG. Ingersoll and DW Moore eds(2005). *Use of Sediment Quality Guidelines and Related Tools for the Assessment of Contaminated Sediments*. SETAC Press, Pensacola, FL, USA, pp. 163–226.

Anderson, BS, Hunt, JW, Phillips, BM et al (2001). Sediment quality in Los Angeles Harbor, USA: a triad assessment. *Environ. Toxicol. Chem.,* 20, 359–370.

Bailer, AJ, Hughes, MR and See, K et al (2002). A pooled response strategy for analyzing multiple responses to develop relative risk rankings in a weight of evidence evaluation of sediment contamination. *Hum. Ecol. Risk Assess.,* 8.

Beyers, DW (1998). Causal inference in environmental impact studies. *J. N. Am. Benthol. Soc.,* 17, 367–373.

Bombardier, M and Blaise, C (2000). Comparative study of the sediment-toxicity index, benthic community metrics and contaminant concentrations. *Water Qual. Res. J. Canada,* 35, 753–780.

Borgmann, U, Norwood, EP, Reynoldson, TB et al (2001). Identifying cause in sediment assessments: bioavailability and the Sediment Quality Triad. *Can. J. Fish. Aquat. Sci.,* 58, 950–960.

Burton, GA, Jr, Chapman, P and Smith, E (2002a). Weight of evidence approaches for assessing ecosystem impairment. *Hum. Ecol. Risk Assess.,* 8: 1657–1673.

Burton, GA, Jr, Batley, GE, Chapman, PM et al (2002b). A weight-of-evidence framework for assessing sediment (or other) contamination: improving certainty in the decision-making process. *Hum. Ecol. Risk Assess.,* 8, 1675–1696.

Burton, GA, Jr, Greenberg, MS, Rowland, CD et al (2005). *In situ* exposures using caged organisms: a multi-compartment approach to detect aquatic toxicity and bioaccumulation. *Environ. Pollut.,* 134, 133–144.

Chapman, PM (1990). The Sediment Quality Triad approach to determining pollution-induced degradation. *Sci. Tot. Environ.,* 97–8, 815–825.

Chapman, PM (1996). Presentation and interpretation of Sediment Quality Triad data. *Ecotoxicology,* 5, 327–339.

Chapman, PM (2000). The Sediment Quality Triad: then, now and tomorrow. *Int. J. Environ. Pollut.,* 13, 1–6.

Chapman, PM, Paine, MD, Arthur, A et al (1996). A triad study of sediment quality associated with a major, relatively untreated marine sewage discharge. *Mar. Pollut. Bull.*,32, 47–64.

Chapman, PM, Anderson, P, Carr, S et al (1997). General guidelines for using the Sediment Quality Triad. *Mar. Pollut. Bull.* 34, 368–372.

Cherry, DS, Currie, RJ, Soucek, DJ et al (2001). An integrative assessment of a watershed impacted by abandoned mined land discharges. *Environ. Pollut.,* 111, 377–388.

Culp, JM, Lowell, RB and Cash, K J (2000). Integrating mesocosm experiments with field and laboratory studies to generate weight-of-evidence risk assessments for large rivers*. Environ. Toxicol. Chem.,* 19, 1167–1173.

De Rosa, CT, Johnson, BL, Fay, M et al (1996). Public health implications of hazardous waste sites: findings, assessment and research. *Food Chem. Toxicol.,* 34, 1131–1138.

ECHA (2010). *Practical Guide: How to Report Weight of Evidence*. ECHA-10-B-05.EN. European Chemicals Agency.

Forbes, VA and Calow, P (2002). Applying weight of evidence in a retrospective ecological risk assessment when quantitative data are limited. *Hum. Ecol. Risk Assess.*

Fox, GA (1991). Practical causal inference for ecoepidemiologists. *J. Toxicol. Environ. Health,* 33, 359–379.

Goodman, SN (1999a). Toward evidence-based medical statistics. 1: The P value fallacy. *Ann. Intern. Med.,* 130, 995–1004.

Goodman, SN (1999b). Toward evidence-based medical statistics. 2: The Bayes factor. *Ann. Intern. Med.,* 130, 1005–1013.

Grapentine, L, Anderson, J, Boyd, D et al (2002). A decision making framework for sediment assessment developed for the Great Lakes. *Hum. Ecol. Risk Assess,* 8, 1641–1655.

Green, RH, Boyd, JM, MacDonald, JS (1993). Relating sets of variables in environmental studies: the Sediment Quality Triad as a paradigm. *Environmetrics,* 4, 439–457.

Hall LW, Jr and Giddings, JM (2000). The need for multiple lines of evidence for predicting site-specific ecological effects. *Hum. Ecol. Risk Assess,* 6, 679–710.

Hawkins, CP, Norris, RH, Gerritsen, J et al (2000). Evaluation of the use of landscape classifications for the prediction of freshwater biota: synthesis and recommendations. *J. N. Am. Benthol. Soc.,* 19, 541–556.

Hill, AB (1965). The environment and disease: association or causation. *Proc. R. Soc. Med.* 58, 295–300.

Kapo, K and Burton, GA, Jr (2006). A Geographic Information Systems-based, weights-of-evidence approach for diagnosing aquatic ecosystem impairment. *Environ. Toxicol. Chem.,* 25, 2237–2249.

Landis, WG and McLaughlin, JF (2000). Design criteria and derivation of indicators for ecological position, direction and risk. *Environ. Toxicol. Chem.,* 19, 1059–1065.

Long, ER and Chapman, PM (1985). A Sediment Quality Triad: measures of sediment contamination, toxicity and infaunal community composition in Puget Sound. *Mar. Pollut. Bull*., 16, 405–415.

Lowell, RB, Culp, JM, and Dube, MG (2000). A weight-of-evidence approach for northern river risk assessment: integrating the effects of multiple stressors. *Environ. Toxicol. Chem.* 19, 1181–1190.

MacDonald, DD, Ingersoll, CG, and Berger, T (2000). Development and evaluation of consensus-based sediment quality guidelines for freshwater ecosystems. *Arch. Environ. Contam. Toxicol.,* 39, 20–31.

Maxon, CL, Barnett, AM, and Diener, DR (1997). Sediment contaminants and biological effects in Southern California: use of a multivariate statistical approach to assess biological impact. *Environ. Toxicol. Chem.,* 16, 775–784.

Mendelsohn, ML, Moore DH, II, and Lohman, PHM (1992). A method for comparing and combining short-term genotoxicity test data: results and interpretation. *Mutation Research,* 266, 43–60.

Menzie, C, Henning, MH, Cura, J et al (1996). Special report of the Massachusetts weight-of-evidence workgroup: a weight-of-evidence approach for evaluating ecological risks. *Hum. Ecol. Risk Assess.* 2, 277–304.

Menzie, C, Burke, AM and Grasso, D et al (2000). An approach for incorporating information on chemical availability in soils into risk assessment and risk-based decision making. *Hum. Ecol. Risk Assess.,* 6, 479–510.

Mumtaz, MM, De Rosa, CT, Groten, J et al (1998). Estimation of toxicity of chemical mixtures through modeling of chemical interactions. *Environ. Health Persp.,* 106, 1353–1360.

National Research Council (2001). *A Risk-Management Strategy for PCB-Contaminated Sediments*. National Academy Press. Washington DC.

Peterman, R (1990). Statistical power analysis can improve fisheries research and management. *Can. J. Fish. Aquat. Sci.,* 47, 2–15.

Peters, C, Parnel, I, Marmorek, D et al (1998). *Conclusions and Recommendations from the PATH Weight of Evidence Workshop*. Vancouver, BC, Canada. *http://www.efw.bpa.gov/Environment/PATH/reports/981008.weight.pdf* .

Pohl, HR, Roney, N, Fay, M et al (1999). Site-specific consultation for a chemical mixture. *Toxicol. Indust. Health,* 15, 470–479.

Posthuma, L, Suter, GW, III, Traas, TP (eds) (2001). *Species Sensitivity Distributions in Ecotoxicology*. Lewis Publishers, Boca Raton, FL, USA.

Read, L, Selly, A, Johns, D et al (2000). Methods for analysis of Sediment Quality Triad data. *Abstr. Soc. Environ. Toxicol. Chem. Annu. Meet.*

Reynoldson, TB, Day, KE, Bailey, RC et al. (1995). Biological guidelines for freshwater sediment based on the BEnthic Assessment of SedimenT (the BEAST) using a multivariate approach for predicting biological state. *Aust. J. Ecol.* 20, 198–219.

Reynoldson, TB and Wright, JF (2000). The reference condition: problems and solutions. In JFWright, JF, Sutcliffe, DWand Furse, MT (eds), *Assessing the Biological*

Quality of Freshwaters: RIVPACS and Other Techniques, pp. 293–303. Freshwater Biological Association, Ambleside, UK.

Smith, EP (2002). BACI design. In El-Shaarawi, AH Piegorsch, WW eds, *Encyclopedia of Environmetrics*, pp. 141–148. John Wiley & Sons, Chichester, UK.

Smith, EP, Lipkovich, I Ye, K (2002). Weight of evidence (WOE): quantitative estimation of probability of impact. *Hum. Ecol. Risk Assess.,* 8, 1585–1596.

Soucek, DJ, Cherry, DS, Currie, RJ et al (2000). Laboratory to field validation in an integrative assessment of an acid mine drainage-impacted watershed. *Environ. Toxicol. Chem.,* 19, 1036–1043.

Suter, GW, II, Norton SB, Cormier, SM (2002). A methodology for inferring the causes of observed impairments in aquatic ecosystems. *Environ. Toxicol. Chem.,* 21, 1101–1111.

Swanson, MB and Socha, AC (eds)(1997). *Chemical Ranking and Scoring: Guidelines for Relative Assessments of Chemicals: Proceedings of the Pellston Workshop on Chemical Ranking and Scoring*. Society of Environmental Toxicology and Chemistry Press, Pensacola, FL, USA.

USEPA (2000). *Stressor Identification Guidance Document*. Office of Water and Office of Research and Development. EPA/822/B-00-025. Washington DC, USA.

Weed, DL (2005). Weight of evidence: a review of concepts and methods. *Risk Analysis,* 25(6), 1545–1557.

Wenning, RJ, Batley, GE, Ingersoll, CG, and Moore, DW (eds) (2005). *Use of Sediment Quality Guidelines and Related Tools for the Assessment of Contaminated Sediments*. SETAC Press, Pensacola, FL, USA. 783 pages.

Wiegers, JK, Feder, HM, Mortensen, LS et al (1998). A regional multiple stressor rank-based ecological risk assessment for the fjord of Port Valdez, AK. *Hum. Ecol. Risk Assess.,* 4, 1125–1173.

Wright, JF (1995). Development and use of a system for predicting the macroinvertebrate fauna in flowing waters. *Aust. J. Ecol.,* 20, 181–197.