FACT SHEET

03

MAY 2016

METALS ENVIRONMENTAL RISK ASSESSMENT GUIDANCE

MERAG

Effects assessment

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

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

In the framework of environmental risk assessment, the main goal is long-term protection of the environmental compartment (water, sediment, soil, and air) under consideration. Typical compartments that are considered for the inland environment and for which environmental threshold values need to be derived are the aquatic (including the sediments), sewage treatment plants (STPs) and terrestrial ecosystems. An environmental threshold value (ETV) for a given compartment (water, soil, sediment) is an environmental effects concentration below which adverse effects on the compartment are not expected to occur. Examples of ETVs are Predicted No Effect Concentrations (PNEC), Environmental Quality Standards (EQS), Water Quality Criteria (WQC), Water Quality Standards, etc. Such values do not consider fate and transport to other compartments, an aspect covered by exposure models. Within this Fact Sheet, the general term ETV will be used unless it concerns specific examples for a certain legislation.

In general, the information presented in this Fact Sheet serves as guidance to collect and evaluate ecotoxicity data with the purpose to derive ETVs for inorganic substances. The target public includes national governmental institutions, industrial users, academics and evaluating experts/consultants faced. The structure of this guidance is the following: Section 2 deals with data quality and data relevance considerations for metals for the selection of hazard data for ETV derivation. Section 3 provides metal-specific guidance on how to aggregate and interpret the selected data and which IT-based tools can be used to derive ETVs depending on the data richness of the metal under scrutiny.

2. DATA COMPILATION AND SELECTION

Ecotoxicity data can be drawn from data required for regulatory purposes as well as from relevant literature and/or internationally recognised databases. Because the data quality of the extracted information may vary considerably between individual source documents, it is very important to evaluate all ecotoxicity data with regard to their overall adequacy for risk assessment purposes. In general, this evaluation involves a review of how well each study was conducted (Section 2.1.3) and how the results are interpreted in order to accept (or reject) a study in accordance with the purpose of the assessment. This adequacy score should cover both the *reliability* of the available data and the *relevance* of the data for environmental risk

assessment/environmental quality setting purposes in general and for metals/metal compounds in particular. These two basic elements are defined as follows:

- *Relevance*: the extent to which data and/or tests are appropriate for a particular hazard identification or risk characterisation.
- *Reliability*: the inherent quality of a test report or publication relating to preferably standardised methodology and the way the experimental procedure and results are described to give evidence of the clarity and plausibility of the findings.

Only those data that can be considered fit for purpose (relevant) and of sufficiently high quality (reliability) should be deemed adequate (acceptable) to be used for the development of environmental quality guidelines/standards.

2.1 Criteria for Data Reliability and Data Relevance

2.1.1 Introduction

Generic: Several scoring systems are available to assess the reliability and relevancy of ecotoxicity data with most of them designed for scoring chemicals in general with no specific focus on metals. The evaluation systems commonly used for studies assessment in different world-wide regulations (eg, EU-ECHA 2008a; EC 2011: USEPA 2004; and OECD 2002- HPV programs) are usually based on the criteria established by Klimisch et al, (1997. However, some studies (Küster et al 2009; Agerstrand et al 2011a,b) indicated the need for an updated evaluation system as the Klimisch system is deemed not sufficiently detailed, tends to favour standard tests and/or tests conducted under Good Laboratory Practices (GLP), and was originally designed for evaluating mammalian toxicity data rather than ecotoxicity data. Recently, a new set of reliability and relevance Criteria for Reporting and Evaluating ecotoxicity Data (CRED-criteria) has been developed (Kase et al 2015a,b; Moermond et al 2015). CRED-aims to improve reproducibility, transparency and consistency of reliability and relevance evaluations of aquatic ecotoxicity studies[1] . The CRED evaluation method provides more detailed guidance on how to evaluate study reliability and relevance, thus leading to greater consistency among individual reviewers (Kase et al 2015b). General data-quality screening recommendations and further reading on those can be found in several guidance documents available on this topic (OECD 1995; EC-TGD 2003; ECHA 2011) and those referenced above.

Metal specificities: These scoring systems have been developed with organic substances in mind. Because metals are naturally occurring substances and the ecotoxicity of metals is

[1] Soil and sediment ecotoxicity data are presently not yet covered by the CRED evaluation system, but the criteria and principles can be applied in the same way.

strongly driven by the amount of metal that is bioavailable, it is imperative to take these specific aspects into account when evaluating the relevance and reliability of toxicity data generated with metals and metals compounds. For example, exposure conditions during the test could be different from those of the natural environment on which the risk assessment is done, eg,the pH or hardness of the test medium may be outside the boundaries of the physico-chemical conditions encountered in a specific environment under investigation or the test organisms could be cultivated under very different conditions and as such be conditioned to a completely different environment.

The subsequent paragraphs highlight some of the more metal-specific issues and new concepts that should be considered when evaluating ecotoxicity data for metals and metals compounds. The main concepts to be covered are:

- Metals are naturally occurring substances and hence natural backgrounds should be taken into account in selecting adequate ecotoxicity data.
- The ecotoxicity of metals is strongly driven by the amount of metal that is bioavailable and this fraction is a function of the physico-chemical conditions of the test media.

They are further described in Sections 2.1.2 on Data relevancy and 2.1.3 on Data reliability

2.1.2 Data relevancy

A first screening is generally performed based on the relevance of the data for the purpose of the aim of the assessment. This is a step that is particularly important for metals/metal compounds as the circumstances under which a metal has been tested could have an effect on its availability for uptake and subsequently its potential to elicit a toxic response.

2.1.2.1 Biological relevancy of the endpoint used

- For risk assessment purposes, both standardised endpoints (survival, growth etc) as non-standardised endpoints (enzyme activity, morphological changes, etc) are used. The ecological relevance of enzymatic assays (eg, phosphatase, urease,…) is more complex and a subject of current debate. Generally, enzyme assays only determine the activity of an individual specific enzyme, whose reaction is frequently a component in a multi-enzymatic process. This enzyme activity in the field might be limited by other factors such as substrate availability or other rate-determining factors. Therefore, the relevance of a single enzyme activity is considered lower than the relevance of net rate measurements of an overall process (Kuperman et al 2014). Moreover, the enzymatic activities are often measured at conditions that are not representative for *in situ* soil/sediment/aquatic conditions for metals and therefore many enzyme assays are of

low relevance for natural exposure conditions. For example, several types of assays are conducted in pH-buffered suspensions and because the metal–enzyme interaction is pH-dependent, this might obscure the relationship with effects in the soil/sediments. For example, tests performed in buffered soil suspensions at a pH value that was greater than 0.5 pH units different from that in the undisturbed soil cannot be considered relevant. Almost all assays use high substrate concentrations (typically several mM), a condition that is unlikely to occur *in situ*. Finally, the colorimetric reaction that is often required in enzymatic assays can also be subject to effects of metals (Nannipieri et al 1997). Therefore, enzyme assays must be assessed with great care, and they are often not selected for the derivation of soil threshold concentrations for metals.

2.1.2.2 Relevancy of the test substance

With the exception of tests conducted using complex materials (like inorganic complex mixtures), tests for metals should generally be conducted with high-purity soluble metal salts if used for the purpose of deriving environmental quality guidelines/standards. Because impurities can have an effect on the toxic properties of the substance under investigation, or have toxic effects themselves, studies involving test substances in which impurity levels are > 0.25% (under limit of classification) should not be used.

2.1.2.3 Relevancy of the test medium

Metals are components of the natural environment and organisms have been evolving in their presence for centuries. Moreover, some metals such as copper and zinc are essential elements, needed for life. For that reason, it is important that metal effects data are obtained under conditions that are relevant for the natural environment. Relevancy relates to both organisms and testing conditions:

- The data used in the effect assessment should ideally be based on organisms and exposure conditions relevant for the site/region under consideration. This could, however, considerably reduce the amount of data to be used. Therefore, data based on test media collected outside the region of interest (eg, on organisms not living in a given environment) also can be used, providing the physico-chemical properties of that other environment are within the range relevant for the environment under study. For the different compartments, the following can be recommended:
    - For soil, only data from observations in natural and/or well-defined artificial soil media (eg, OECD soil) are considered as relevant for the terrestrial effect assessment. Tests performed in other substrates (eg, nutrient solution (hydroponics), agar, pure quartz sand, and farmyard manure) are judged as not

representative for exposure in natural soils, and hence their relevance for risk assessment is considered low.

- For sediment, only data from observations in natural and/or well-defined artificial sediments are considered as relevant for the sediment-effect assessment. Tests performed in other substrates (water only, pure quartz sand, shredded paper towel) are judged as not representative for exposure in natural sediments, and hence their relevance is considered low. In cases where no data are available in natural or artificial sediments, a lower tier Kd-based (partition coefficient) extrapolation might be suitable for screening purposes.
- For all compartments, only those test media should be used that are within the physico-chemical boundaries of the environment under study. This relates to physico-chemical conditions eg, pH, water hardness, dissolved organic carbon (DOC), total organic carbon (TOC), grain size, temperature etc as well as background metal concentrations. The latter factor may be of particular importance, eg, the concentration of essential elements during the test should be within the organisms' homeostatic range for that element, to avoid, notably, deficiency effects.

- The fact that metal/metal compounds are naturally occurring substances should be taken into account when selecting toxicity data because phenomena such as acclimatisation and adaptation are of importance. Indeed, due to the ubiquitous presence of metals in the natural environment, organisms have become conditioned to these backgrounds because they have evolved in the presence of the natural metal background concentrations. For this reason, exposure of organisms to the natural background level reflects in fact the theoretical lower limit of the predicted no effect concentration (PNEC) ie, a concentration, which from an evolutionary perspective, does not present a potential disruption of the genetic pool composition of a species. This theory is applicable for all metals and is even more crucial for essential metals[2]. As a result, the sensitivity of organisms to metals is determined to a large extent by the bioavailable concentration that the organism experienced before testing and their developed capability to cope with this concentration. Moreover, organisms cultured in media with a low essential metal concentration[3] may also exhibit an overall decreased fitness (deficiency issues) and become more sensitive to stress, including exposure to

[2] An element is considered essential when (1) it is present in living matter; (2) it is able to interact with living systems; (3) a deficiency results in a reduction of a biological function, preventable or reversible by physiological amounts of the element (Mertz, 1974).

[3] This is especially the case in artificial media, because these media contain no or very little (essential) micronutrients.

metals, even essential ones. A good example, and one whose implications pervade the metals toxicity literature, involves interpretation of test data from laboratories that used waters of the upper Great Lakes, which have naturally low metals concentrations due to the characteristics of their watersheds, and also low natural concentrations of major ions. For example, in the EU risk assessment for zinc, those ecotoxicity tests conducted with Lake Superior water with Zn concentrations below 1 µg/L were excluded from the database (EU Zn RAR 2006).

Although local organisms are usually cultured in Great Lake waters that have been "salted up" to "moderately hard" concentrations of major ions, and tests conducted in moderately hard waters, the organisms have never experienced any transition metal exposure conditions other than their metal-poor natural waters (Van Assche et al 1996, 1997). Conversely, organisms cultured in media with elevated metal concentrations (both essential and non-essential metals, eg, natural waters or contaminated waters) may become less sensitive to those metals. This phenomenon is related to the "biogeochemical region" concept (Fairbrother and McLaughlin 2002). A good example, from nature, of the ability of a nominally metal-sensitive organism to adapt (or perhaps acclimate) to a high-metal biogeochemical region is populations of brown trout in the River Hayle in Cornwall, England (Durrant et al 2011; Uren Webster et al 2013). Brown trout in this river with naturally and anthropogenically elevated metal levels (for millennia, going back at least to Roman mine works) are able to thrive at concentrations that are up to ten times the UK EQS values, and have some of the highest tissue body burdens ever measured. The fact that brown trout are successfully reproducing in these metal-rich waters (see Table 1) with metal concentrations far exceeding copper and zinc concentrations 12 µg Cu/L and 55µg Zn/L) that brown trout are known to avoid (Woodward et al 1995) suggests that this may be a genetic adaptation, because robust metal regulation is active from the egg stage on.

In order to evaluate the extent of a successful adaptation in a high-metal biogeochemical region, evidence should be provided to ensure that impairment (eg, reproductive effects) has not occurred in comparison with other populations thriving in less metal-rich waters.

Region	Dissolved Cu and Zn concentrations (µg/L)	Trout population present
Upper region (low metal contamination)	Cu: 4; Zn: 28	+
Middle region (highly contaminated)	Cu: 94; Zn: 760	+
Lower region (moderately contaminated)	Cu: 35; Zn: 530	+

Table 1: *Dissolved copper and zinc concentrations between 1997-2003 in the river Hayle (Cornwall, south-west England)*

- Ideally, testing for ETV derivation should only use those data sets where background concentrations in the culture medium (ideally both essential as well as non-essential metals) are similar to the clearly defined, relevant conditions of the biogeochemical region[4] under investigation and are also representative for natural conditions suitable for the organism. However, it is acknowledged that this type of information is not always reported and hence not always useful as a selection criterion. If the information is available (occurring especially for the major metals), the information can be used to consider whether or not to use test results where the organisms were cultured under natural background conditions that deviate from the conditions encountered in the environmental compartment under consideration. In particular, it is a prerequisite that the essential metal concentration in the culture medium should be at least equal to the minimal concentration not causing deficiency for the test species used, ensuring that the effect of the substance is measured and not the one from the culture conditions. In addition, it should be preferably in the range of the natural background concentrations for the area under consideration (US, EU lowland, or Nordic shield,…) For example, the FOREGS[5] database provides some information on natural background concentrations in water and sediments (Salminen et al 2005). Concentrations of non-essential metals should fall within the natural background variation of these metals.

[4] The biogeochemical-region approach arises from the fact that different eco-regions can be identified based upon climatic factors, latitude, and elevation. Within eco-regions, sub-eco-region (also called biogeochemical-regions) can be differentiated based upon the natural background concentration (see also Reimann and Garret 2005 on the means to determine background concentrations) of the metal under consideration and the presence of well-defined abiotic factors that influence metal bioavailability.

[5] http://weppi.gtk.fi/publ/foregsatlas/

- Defining minimal levels of metal background for selection of relevant culture media should only be performed in case there is scientific evidence that acclimation/adaptation phenomena are relevant for the metal under investigation. If no direct information is available on the background concentrations of the metals in the culture medium, second-line evidence (eg, metal concentrations in river water used for maintaining the cultures could have been measured in other studies) and expert judgment can be used to support any decision taken on this issue. If background concentrations have been reported and test organisms have been cultured in conditions that are outside the natural background concentration ranges, such data should be carefully evaluated and be discarded or only used as supportive information depending on expert judgment. It is, however, recognised that this may lead to a reduction in the number of useful ecotoxicity data points which may even sometimes limit the possibility of using a Species Sensitivity Distribution (SSD).

2.1.2.4 Relevancy of the test species

- For preliminary, screening-level risk assessment purposes for undefined or unspecified geographical regions (eg, for chemicals management assessments) in general both endemic and non-endemic species are usually considered relevant.

- In addition, the species assemblage should focus on representative species for the different compartments, eg,, primary producers (higher plants & algae), primary consumers (invertebrates) and secondary consumers (fish, amphibians) for the aquatic ecosystem; bacteria and ciliates for STPs; sediment-dwelling organisms with different exposure routes, feeding habits and ecological niche for the sediment compartment; primary producers (plants), consumers (invertebrates) and decomposers (microbial mediated processes) for the terrestrial environment.

- The relevancy of the trophic levels and endpoints to be considered in the effect assessment may differ in breadth and detail among legislations. For example, while certain algae and plant species are sensitive to metals, different approaches are used in terms of incorporating plant/algae data. Some jurisdictions combine them into the same SSD (eg, REACH/Europe or Japan), whereas other jurisdictions (eg, USEPA) keep plant and algae data separate from combined invertebrates/fish databases. Regardless of the way they are treated, the recommendation is that data of all relevant taxa (including algae/plant data) need to be collected.

- If a risk assessment is **conducted for a certain region** (eg, eco-region) or local site, the selection of ecotoxicity data should take into account the relation between 1) test conditions, 2) area/region, and 3) test species (Figure 1). It is recommended to select or conduct testing with a test organisms that is relevant for the area under review. Moreover, an effect data point should only be accepted when the test conditions (ranges of metal background concentrations, abiotic factors such as pH, Ca, Mg etc) are relevant for both organism and area. Until now, these issues did not receive a lot of attention, but it will become more important as data will be further extrapolated to other climate zones or agro-ecological regions. If bioavailability models are available to correct (eg normalise) for differences in abiotic factors, tests with deviating test conditions could be used.

- If a particular sensitive species should be removed in applying these relevance criteria, then some caution is needed. In case the species is closely related to a known regionally relevant species, if there is a paucity of data for other species, or if the species may have existed in the region until recently (potentially due to anthropogenic activity: eg, past diffuse sources), a precautionary approach could be considered by retaining the data point.

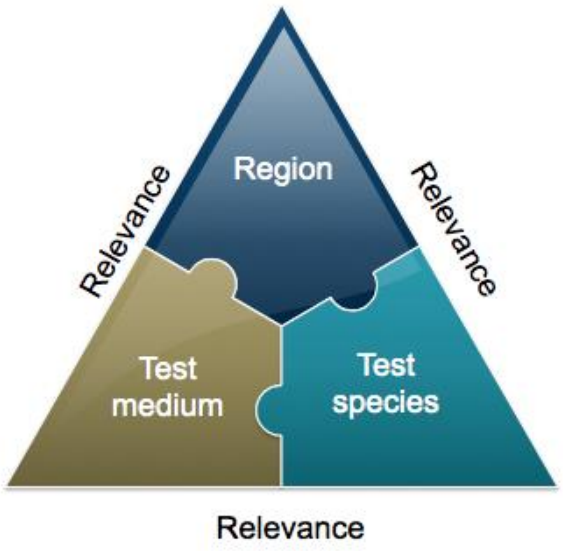

Figure 1: *A region-specific assessment should take into account region-specific information with regard to the relevance of the test medium and test species for the region under scrutiny*

- With notable exceptions (eg, the zebrafish *Danio rerio*), most effects data sets on metals were conducted with species and/or conditions relevant for the temperate climate zone. Currently, some research has been conducted to evaluate whether temperate species

are more or less sensitive than tropical or cold area species to check the relevance of existing data sets for these other regions. As metabolic rates generally increase with increasing temperature, it could be expected that tropical species would be more sensitive. However, this need not to be the case because increase in temperature may also increase rates of detoxification counteracting the previous phenomenon. Chapman et al (2006) compared acute marine toxicity data and could not discern a clear pattern in sensitivity differences. Wang et al (2014a) also showed that there was no significant difference between Chinese and American taxa and, in a similar study, differences between temperate and tropical salt water species, acute sensitivity appeared to be small (Wang et al 2014b) with no clear indication of increased sensitivity. Based on the present evidence for metals there does not seem to be a real difference in sensitivity, so therefore, existing data can be used. It could be recommended to add a validation check with an endemic tropical or cold area species relevant for the region under consideration.

2.1.2.5 Relevancy of exposure duration

- Both acute and chronic data can be used for the derivation of ETVs. Preference should be given to the use of chronic data if available, which is for metals often the case. Acute toxicity tests are often only knock-out studies with limited ecological relevance. However, they are still useful with regard to certain peak exposures (eg, intermittent releases, calamities), deriving short-term environmental quality guidelines and its use for classification purposes. Also, for data-poor substances, they can be useful to derive a chronic ETV when using an assessment factor approach. Finally, as many substances have different mode of actions at short-term and long-term exposures, it is useful to consider both types of studies. The 72-h algal growth inhibition test is a chronic test but the EC_{50} is treated as an acute value, for example, for classification purposes. Following the latest OECD requirements, relatively short-term studies, focusing on sensitive life stages rather than focusing on the full life stage are also deemed chronic studies (eg, root elongation assay for terrestrial plants, ISO 11269-1 (ISO, 1993).

- When there is a lack of chronic data for one or multiple trophic levels, it may be possible to use acute data in combination with appropriate acute to chronic ratios (A/C). A/C ratios are relatively small and usually smaller than a factor 10 (Yungsong et al 2014). Quantitative ion character-activity relationships (QICARs) or quantitative cationic-activity relationships (QCARs) could be used in the complete absence of experimental data (Ownby and Newman 2003; Walker et al 2003) as is the case for some data-poor inorganic substances. However, more research efforts are needed in this field to develop

and validate appropriate models. If no appropriate models are available, the ETV has to be derived from chronic data.

2.1.3 Data reliability

All relevant data selected must be further screened for the inherent quality of the results. As indicated in section 2.1.1 different scoring systems are available (Klimisch or CRED). Whatever, evaluation system-used study data are typically categorised into different reliability classes:

- *Reliable without restrictions (Reliability class 1);*
- *Reliable with restrictions (Reliability class 2);*
- *Not reliable (Reliability class 3);*
- *Not assignable (Reliability class 4).*

A checklist for evaluating the general quality of ecotoxicity studies is provided in Table 2. These criteria are mostly not metal-specific, they simply adhere to the principles of good study conduct.

Test protocol • standard method available • test performed under GLP[5]
Type of test • standard test or non-standard test • endpoint used • test duration • test conditions (eg, static or flow through)
Description of test material and methods • test set-up, measuring chamber/device • test material/method (including purity), spiking method, dilution water if applicable, equilibration period prior to exposure/testing • test organism, including size (age), origin, number of organisms per replicate • test design (# replicates that should be used) • type of food given (chronic tests), test metal concentrations in food
Description of physico-chemical properties of the test media • proper description and control of physico-chemical conditions (eg, pH, • Table) that may affect the speciation (ie, availability) and bioavailability and toxicity of the metal studied and hence may influence the outcome of a test
Chemical analysis

• evidence is given that concentrations were maintained during the test (< 30% variation)
  - test metal concentrations during the test are measured
  - test metal concentrations are not measured, but indication is given that the nominal concentrations are close to actual concentrations
• for sparingly soluble metals, measured data on the dissolved fraction are always required in order to obtain reliable toxicity test data

Concentration-effect relationship • acceptable control response (mortality, reproduction, growth, etc) • appropriate statistics used, 95% confidence limits reported or data on the relationship given amenable to further analysis to derive a suitable $L(E)C_x$ value • concentration range is given • at least 2 different concentrations must have been tested besides the control • a concentration-related response should be clear (a progressive effect should be observed as a function of the dose) • hormesis effect observed or not

Table 2: *Non-limitative checklist of criteria for the evaluation of the reliability of ecotoxicity studies used for risk assessment and/or ETV setting*

These criteria are outlined in more detail further, together with some **more metal-specific focus** points.

2.1.3.1 Type of test

- **Environmental hazard classification** is based on the comparative toxicity at equal conditions between chemicals compared to a reference criteria scheme. It is in such cases highly recommended to use preferentially selected data on standard (aquatic) test organisms to ensure focusing the difference in hazard categorisation based on the difference of intrinsic toxicity of the metals rather than between standard and non-standard species.

- Both standard test organisms and non-standard species can be used in the framework of **a risk assessment.** In general, toxicity data generated from standardised tests, as prescribed by organisations such as OECD, ISO and USEPA will need less scrutiny than non-standardised test data, which will require a more thorough check on their

compliance with quality criteria before being used. GLP[6] and non-GLP tests can be used provided that the latter fulfill the stipulated requirements.

- In the aquatic environment, both static and semi-static tests (renewal) and flow-through tests can be used. The results of the latter should be handled with care for metals/metal compounds and it should be evaluated if enough equilibration time was provided to allow for equilibrium partitioning of the bioavailable metal fraction. For sediment testing, semi static and flow-through test designs are preferred because the sediment may act as a source of dissolved metals to the overlying water column, (Wang et al 2004). This could result in a build up of metals in the overlying water causing toxicity via this route. If the latter is observed, the tests should not be considered reliable. Static test designs are more prone to this phenomenon than test systems with sufficient water renewals.

- For sediment and soil testing, adequate time should elapse between mixing metal or metal compounds into the test medium and introducing biota (plants or soil/sediment species). Initial partitioning of metals, taking place within hours or days after addition of soluble metals to a moist soil, is often followed by much slower reactions, termed fixation or ageing, that further decrease the bioavailability of added metal with time (eg, Buekers et al 2008; Ma et al 2006a,b). However, equilibration in a water-only system will be reached within hours or days, while for sediments and soils full chemical equilibrium may only be obtained after several months or even years. Both short equilibration times and high spiked metal concentrations in sediments/soils will accentuate partitioning of metals disproportionately to the dissolved phase (Lee et al 2004; Simpson et al 2004). Results from standard tests in freshly spiked soil also generally overestimate toxicity in realistic field conditions (eg,, Lock et al 2006; Ooorts et al 2006; van Gestel et al 2012). Because it is practically not feasible to equilibrate all soils for several months after spiking with a soluble metal salt, correction factors are derived for the effect of this slow equilibration of metals in soil on their toxicity (see Section 3.3.1.1 lab-to-field factor). It is, however, recommended that all soils are equilibrated for approximately one week after mixing the metal into the soil before introducing the test species in the soil and starting the ecotoxicity assays.

- Precautionary steps are needed to achieve more stable and environmentally realistic partitioning of metals in spiked sediments. Simpson et al (2004) investigated the equilibration and bioavailability of metals in laboratory-contaminated sediments in order

[6] GLP = Good Laboratory Practice

to provide better guidance on acceptable procedures for spiking sediment with metals for use in the development of whole sediment toxicity testing. It was demonstrated that sediment spiked with Ni required a relatively long time for equilibration – as long as 70 days (d), compared to 15 d for copper, 40 d for zinc, and 45 d for cadmium. The addition of metals to the sediment causes major decreases in pH and an increase in redox potential. Based on these and other findings, new spiking methods have been developed and applied in sediment toxicity tests with the purpose to reduce the diffusional loss of nickel from the sediment phase into the overlying water in laboratory sediment toxicity tests (Brumbaugh et al 2013; Besser et al 2013). In addition to spiking-method considerations, procedures for sediment toxicity testing were also modified to ensure environmentally realistic partitioning of metals among sediment, porewater, and overlying water (Brumbaugh et al 2013). Metal concentrations in the overlying water of sediment toxicity tests should be measured for the test substance, and testing should be initiated only when overlying water concentration reaches acceptable levels, eg, EC_{10} levels derived from water only tests. Simulated aging and weathering processes may also be desirable but currently are not embedded in standard sediment test protocols.

2.1.3.2 Description of test material and methods

A detailed description of methods employed in the study should be provided. This description should include at least the method of test medium preparation, time of spiking, recorded observations. To calculate free ion concentrations with speciation codes, the concentrations of dissolved major anion and cations, Fe, Mn, Al, dissolved organic carbon (DOC), pH are required (see Fact Sheet 5 for further information on which speciation models could be used).

Furthermore, the organisms used should be uniform in size and age and represent a sensitive life stage. The test results should allow a proper statistical analysis and the experimental design should provide sufficient replicates per test concentration to derive a high-quality NOEC value or alternatively less replicates but more concentration levels to derive a high-quality $L(E)C_x$ value[7].

2.1.3.3 Description of physico-chemical test conditions

Considering the strong influence of water physico-chemistry on metal toxicity, the physico-chemical conditions (metal concentrations abiotic factors and biotic factors) should be

[7] $L(E)C_x$ = the concentration that causes x % change in response (e.g mortality, immobility) during a specified time interval. NOEC = No Observed Effect Concentration is defined as the test concentration below the lowest concentration that did result in a significant effect (LOEC = Lowest Observed Effect Concentration) in the specific experiment.

adequately described and reported to allow for carrying out corrections for bioavailability (see Fact Sheet 5). The aquatic medium should be characterised by DOC concentration, hardness, pH, alkalinity and any other specific parameter of importance to the metal in question. In Table 3 an overview is given of physico-chemical characteristics for each compartment that should preferably be reported and fall within the tolerance limits of the test organisms. If these limits are exceeded, the test has to be considered not reliable and receive a Klimisch 3 score.

Water	**Sediment**	**Soil**
Required		
Total and dissolved Me concentrations[8] pH Hardness (Ca^{2+}, Mg^{2+}) Dissolved Organic Carbon (DOC)	Total metal concentration Total organic carbon (TOC) Acid Volatile Sulfides (AVS) Fe/Mn oxides	Total metal concentration Total organic carbon (TOC) pH Cation Exchange Capacity (CEC) Fe/Mn Oxides
Supportive information		
Other inorganic ligands (SO42-, Cl-, Na +, K, +, etc) Culture conditions: metal background concentration, level of essential metals in case of testing essential metals.	Particle size (sand, silt and clay content) Pore water chemistry (total and dissolved metal concentration, DOC, hardness, conductivity/salinity, ammonium etc) Culture conditions: metal background concentrations	Particle size (sand, silt and clay content) Culture conditions: metal background concentrations

Table 3: *Recommended physico-chemical parameters per compartment that should be reported when performing toxicity tests with metals and metal compounds*

- Test media containing chelators (eg EDTA or high DOC levels) should be avoided as they may decrease the metal bioavailability and toxicity.

[8] Different definitions for the dissolved fraction exist. Most often the dissolved fraction in ecotoxicity tests refers to the fraction that passes through a filter of 0.45 µm. It should be noted, however, that this definition may not necessarily refer to the metals in solution. In the range of 0.01-0.45 µm, colloid inert particles that remain suspended may exist and these could account for 50 % or more of the "dissolved" 0.45 µm fraction.

- A sufficient long equilibration time should be respected before the organisms are introduced in the medium.
- With regard to metal concentrations, the current state-of-the-science dictates that only test results where metal concentrations have been measured should be used. Absence of measured exposure concentration data (nominal results) is a clear reason to reject a data point.

2.1.3.4 Chemical analysis

- With regard to metal concentrations, the current state-of-the-science dictates that only test results where metal concentrations have been measured should be used. Absence of measured exposure concentration data is a clear reason to reject a data point. Analytical measurements of the metal concentrations in the test solution allow to (1) exclude human error related to the preparation/addition of test substance solutions; (2) because metals are natural elements, it is therefore important to know the total metal concentrations organisms are exposed to, including the metal background levels in the control/dilution test medium. For example, in river waters and soils, the metal levels in controls can already be relatively elevated in comparison to the metal added as test solution. In this respect, it is important to also consider that organisms adapt to the culture media, not test media. Only studies with measured metal concentrations receive a reliability score 1.
- Measured data should reflect the dissolved fraction (< 0.45 µm) and/or total metal concentrations. Measurements of dissolved metal concentrations are critical to the assessment of sparingly soluble metals[9] (particles and precipitation may occur) and in the use of natural waters as test media (adsorption to suspended solids may occur). A description of the filter methodology and its efficiency are therefore most relevant.
- As for all chemicals, if the solubility product of a metal in water is exceeded, the test has to be considered as unreliable. Results from tests where a visual precipitation is observed should be discarded. However, because inspection by the eye for precipitation is rather subjective, it is recommended to analytically determine both the total and the dissolved metal concentration over time as precipitation kinetics may take hours and sometimes days. The absence of a visual precipitation does, however, not exclude that

[9] Different definitions for the dissolved fraction exist. Most often the dissolved fraction in ecotoxicity tests refers to the fraction that passes through a filter of 0.45 µm. It should be noted, however, that this definition may not necessarily refer to the metals in solution. In the range of 0.01-0.45 µm, colloid inert particles that remain suspended may exist and these could account for 50 % or more of the “dissolved” 0.45 µm fraction.

sometimes colloids may still be present that could affect the test results (eg, through oral toxicity). Colloids should not be considered as inert particles because they can be very dynamic.

- If it is not mentioned whether the reported toxicity values are based on measured or nominal concentrations, they should be considered as nominal concentrations and should not be used achieving a reliability score 3.

2.1.3.5 Concentration-effect relationships

- It is recommended that the L(E)C_x value should not be extrapolated below the lowest applied (non-zero) concentration. According to Reily et al (2003) and ISO (ISO 2006), estimation of L(E)C_x values outside the concentration range tested introduces a great deal of uncertainty. If the resulting L(E)C_x value should be below the lowest applied control level (background level) or essentiality level, its reliability/relevance has to be questioned (another confounding factor in this respect is the hormesis phenomenon which for essential metals can be very important - hormesis is a term used for generally favorable biological responses to low exposures to toxins or other stressors).

- **Care should be taken in evaluating non-monotonic dose-response curves (NMDRCs).** For metals: Non-monotonic dose response relationships show biphasic or bidirectional responses to dose appearing in U-shaped or inverse U-shaped graphic forms. For essential elements, in theory, harmful effects may be observed due to metal deficiency caused by very low metal concentrations caused by particular bioavailability conditions used in the test. Therefore, essential metals quite often exhibit biphasic dose-response curves visualized as U shaped or inverse U-shaped graphic forms. The concept that many metals are required for organisms health at one range of concentrations and are toxic in quantities that may be either more or less than that range has been referred to as the “window of essentiality” (Hopkin 1989) or the Optimal Concentration range for Essential Metals (OCEM), (Alloway 1995; Fairbrother and Kapustka 1997; Van Assche et al 1997). But also for some non-essential metals that are present at low doses an increased performance in, for example, growth, reproduction at low metal doses can be observed. This phenomenon is often referred to as the hormesis effect. While hormesis should be a well-accepted phenomenon in ecotoxicology (Calabrese 2005), its widespread inclusion in applications of ecotoxicology, such as environmental risk assessment and calculation of environmental quality guidelines/standards, represent a substantial scientific challenge (Kefford et al 2008; Hoffman 2009). For example, in such cases, the conventional log-logistic dose-response model is less useful to be used to fit the toxicity data, and adaptations need to

be made. For example, the linear-logistic model of Brain and Cousens (1989) has been extended in the case of hormesis to allow EC_{50} and EC_{10} calculations (Van Ewijk and Hoekstra 1993; Schabenberger et al 1999; Cedergreen et al 2005).

2.1.4 Conclusion

Only ecotoxicity data on metals that comply with the above-mentioned criteria for reliability and relevance can be considered valid for risk assessment purposes for metals/metal compounds. However, the evaluation of toxicological data should not follow a rigidly fixed format, but rather should be used in a flexible manner using expert judgment[10]. The evaluation could also allow for special consideration on a case-by-case basis and may vary depending on the aim of the risk assessment (eg, consideration of endemic species). Minimal data requirements for getting an overall adequacy score of 1 or 2 are: measured test concentrations, proper description of test system, test setup (doses, replicates), proper statistics, acceptable test performance, and representativeness of the test organisms and bioavailability parameters for the environmental compartment under investigation. For older data, some of the above-required information could be lacking and in case of data-poor metals a bit more flexibility can be used regarding the information reported. For transparency reasons, it should be clearly documented which studies are being rejected and on what ground.

3. DERIVATION OF ENVIRONMENTAL THRESHOLD VALUES (ETV)

Within the different type of regulations dealing with setting environmental quality guidelines/standards, different methods are in place to derive thresholds below which unacceptable effects are unlikely to occur. For example , used in a preventive way, EC10 values are selected for the purpose of setting Canadian Water Quality Guidelines (CCME, 2007). Canadian Soil Quality Guidelines are used in a remedial way and hence have different methods and preferred endpoints ($L(E)C_{25\text{-}50}$) to set remedial goals. Typical compartments that are considered are for the inland environment aquatic (including the sediments and terrestrial ecosystems).

[10] For example, information on metal concentrations in the culture medium will not always be available. In those cases, the toxicity data of studies lacking this information could still be used when no other information is available.

3.1 Use of L(E)Cx Data

Different statistical estimates (eg ECx, LCx, NOEC, LOEC, MATC, NEC etc) are in use to express the results of ecotoxicity tests (OECD 2006) but a certain hierarchy in the use of certain statistical estimates can be discerned when using them to derive an ETV. Overall, there is a recommendation to give more weight on concentration-response modeling (eg, regression methods ECx) instead of hypothesis-testing methods (NOEC/LOEC) with the use of EC_{10} as the preferred endpoint for deriving safe thresholds. However, as NOEC data are in some cases prevalent, it is recognised in some jurisdictions that a combination of NEC, ECx and NOEC values will, for some time, be unavoidable (Batley et al 2014). Different statistical estimates can also be used depending on the purpose of the assessment. For example, permit setting and derivation of national/regional safe thresholds quite often have the objective to protect 95% of the species and are using low effect concentrations. Other estimates can be used for site clean-up determination, impact assessment, natural resources damage assessment, or life cycle assessment (Table 4).

Statistical endpoint	Purpose
$L(E)C_{10}$ (preferred)	Derivation EQS (EU), Derivation Canadian Water Quality criteria (CCME 2007) Impact assessment MoE Japan MoE Korea
No Observed Effect Concentration (NOEC)	
EC_{20} (preferred)	Derivation Water Quality Criteria (USA) (USEPA 1985; USEPA 2009)
Maximum Allowable Toxicant Concentration MATC (= geomean of NOEC/LOEC)	Derivation Water Quality Criteria (USA) (USEPA 1985; US-EPA 2009)
EC_{10}, NEC, NOEC	Derivation Australian and New Zealand Water Quality Guidelines (Batley 2014)
$L(E)C_{50}$	Life cycle assessment Derivation short-term water quality criteria (USA)
$L(E)C_{25\text{-}50}$	Site clean-up determination (eg Canadian Soil Quality Guidelines)

Table 4: *Overview of the use of different statistical estimates depending on the purpose*

The selection of the effect level should not be made independent of the overall method to derive threshold concentrations for a metal in an environmental compartment (Section 3.3). For example, different combinations of effect level (EC_x) and protection level (HC_p) are considered for the derivation of soil clean-up values from a species sensitivity distribution (SSD) (Checkai et al 2014). The question arises about which HC_p level of protection is most appropriate for each type of EC_x SSD: is a higher number of species affected reasonable when the adverse effect per species is relatively small, or is it better to have a larger effect on fewer species?

3.2 Aggregation/Selection of L(E)Cx and NOEC Data

In small data sets (eg, for data-poor metals), most often there is preference for selecting the most sensitive value rather than using a geomean value, eg, the lowest reliable E(L)C_x or NOEC value available (CCME 2007). The use of the lowest value provides a precautionary approach, especially when a wide variation is observed between the lowest and highest data points for a given species. However, it should be realised that some of the lower toxicity values reported in literature may be the results of poor organism health, operational conditions, or may just reflect differences in abiotic test conditions (bioavailability), and may therefore not reflect the intrinsic sensitivity of the organisms to a given toxicant. In those cases, the lowest value may be omitted from the database.

For data-rich substances such as often the case for metals/metal compounds, multiple data points can be available from reliable studies for a given species and endpoint. These results will be subject to several sources of variability, such as differences in physico-chemical characteristics of the test media, which can affect metal speciation and bioavailability, inter- and intra-laboratory variability, as well as inherent intra-specific heterogeneity in test organism sensitivity. The most straightforward way to handle situations in which multiple data points exist for a given test species and endpoint, is to use the lowest value. When it is apparent from the data that the observed differences in test results for one species are due to differences in bioavailability in the test media, then the use of the lowest toxicity value should be avoided. Instead, it is recommended that bioavailability normalisations should be conducted prior to appropriate aggregation (grouping) and this according to the best level of scientific knowledge available (eg organic carbon normalisation, hardness correction, bioavailability models). In case fully developed bioavailability models are available (eg, Biotic Ligand Models), the scope of the data gathering can be broadened (provided that the models are validated over a broader range of conditions). Guidance on the principles and how bioavailability can be incorporated are given in Fact Sheet 5.

When it is apparent from the data that the observed intra-species variability in toxicity test results can be assigned to differences in bioavailability and no bioavailability model is available to normalise the data, the effect data should be grouped by similar ranges of abiotic factors that control the bioavailability of metals. This grouping should preferentially be conducted such that it reflects the range of abiotic factors encountered in the region under evaluation (eg, soft water scenario). It is recommended to define region-specific boundaries of these physico-chemical parameters for the selection of relevant test media. In this regard, both natural and artificial test media are acceptable, provided that major physico-chemical characteristics that alter the bioavailability of the metal (ie, pH, major cations, anions, DOC for the water compartment, pH, effective CEC, organic carbon content and clay content for the soil compartment or AVS and information on organic content for the sediment compartment) are similar to the range of the physico-chemical conditions encountered in the waters, soils, or sediments under investigation.
A further potential for refinement is in the case if acclimation/adaptation is important. Then it is worthwhile if test results should be grouped on the basis of the similarity of the background in the culture medium in relation to the background of the environment under evaluation. This type of grouping should be done in cases where the biogeochemical region concept is relevant and can be applied. In this concept, it is recognised that background concentrations of a metal in a given region can differ between ecosystems, resulting in different sensitivities to the toxic effects of metals due to acclimation or adaptation. In this way, effect data sets can be divided into different biogeochemical region groups. Typical biogeochemical regions for the metal under consideration should be based upon clearly distinguished ranges of natural background concentrations and ETVs should be derived for each of these biogeochemical regions.

Summary of grouping rules of selected data

In general, the following grouping rules can be applied:

- If for one species more than one L(E)C_x or NOEC value based on the same toxicological endpoint and tested in a medium with the same physico-chemical properties is available, these values are averaged by calculating the geometric mean, resulting in the “species mean” L(E)C_x value. In case of a flawed data set: eg, only two data points are available and one represents a very low value and another a high value, it is recommended to repeat testing and take the geometric mean of all data.
- If for one species several L(E)C_x values based on different toxicological endpoints are available, the lowest value is selected. The lowest value is again determined on the basis of the geometric mean if more than one value for the same endpoint is available.
- In some cases, L(E)C_x values for different life stages of a specific organism are reported in the same study. If from these data it becomes evident that a distinct life stage is more

sensitive, the result for the most sensitive life stage is selected. The life stage of the organisms is to be indicated in the tables as the life stage at the start of the test (eg, fish: yearlings) or as the life stage(s) during the test (e,g eggs → larvae, which is a test including both the egg and larval stages).

- If acclimation/adaptation is important, test results should be grouped on the basis of the similarity of the background in relation to the culture medium with the background of the environment under evaluation.
- When it is apparent from the data that the observed intra-species variability in toxicity test results can be assigned to differences in bioavailability, results should be normalised using appropriate bioavailability models (eg, Biotic Ligand Models, see Fact Sheet 5) towards the same abiotic conditions prior to calculating a species mean value based on all data for the same species. For example, for nickel only reliable toxicity data from tests within the boundaries of the developed BLMs have been used for establishment of the Predicted No Effect Concentration (PNEC) (Table 5). Definitions of the relevant environmental conditions and the exclusion of otherwise reliable ecotoxicity data relative to these conditions may need to be adapted for other regions.

Table 5: Ranges of pH and hardness used for data selection (SIDS Ni 2008)

Test organisms	pH range	Hardness range (mg/L $CaCO_3$)
Algae- *P-subcapitata*	5.7-8.2	20-480
Higher plants- *H. vulgare*	4.1-7.5	NA
Invertebrates- *D. magna*	5.9-8.2	6-320
Invertebrates-*C.dubia*	6.5-8.2	6-320
Fish-*O. mykiss*	5.4-8.5	20-310

In case no bioavailability model is available to normalise the data, but qualitative information is available on the abiotic parameters controlling the bioavailability, the effect data should be grouped by similar ranges of abiotic factors that control the bioavailability of metals. The grouping should preferentially be conducted such that they reflect the range of abiotic factors encountered in the region under evaluation (eg, soft water scenario). In case toxicity results for the same species vary strongly for tests performed in test media covering a wide range in abiotic parameters (eg, pH), but no information is available on bioavailability corrections of the metal, it is not recommended to group results by calculating a species mean as the geomean of the results for the most sensitive endpoint because this value may not be conservative as it will not be protective for the most sensitive water/sediment/soil environment. In this case, it is recommended

to exclude those test results where metal toxicity is being mitigated to the presence of abiotic factors interfering with the bioavailability of the metal.

3.3 Approach for the Derivation of ETVs

3.3.1 Introduction

Environmental Toxicity Values (ETV) such as Predicted No Effect Concentrations (PNECs), Environmental Quality Standards (EQS etc) can be derived for different environmental compartments: aquatic, sediment, terrestrial, microbial activity in STPs etc. These values are calculated according to the following approaches:

- When data are available for a limited number of species (data poor substances), ETV setting is based on the use of assessment factors reflecting the degree of uncertainty in extrapolating from laboratory toxicity test data for a limited number of species,
- If ecotoxicological data are available for a sufficient number of species (data-rich substances), the use of a statistical extrapolation method is recommended.

In any case, the relevance of the ETV derived should be evaluated. For example, a reality check should be conducted to evaluate if the ETV is below or above the natural background of the metal under consideration or in case of essential metals if the ETV is not situated in the deficiency levels. If the ETV is below or close to background/essentiality levels due to use of assessment factors, there is a need for additional data testing.

A pragmatic way to deal with different backgrounds is to use an added approach. This approach has been further developed in Annex 1 of Fact Sheet 1.

3.3.2 Calculation of ETV using assessment factors (data-poor metals)

For some metals/metal compounds, the amount of data available for predicting ecosystem effects will be limited. In these circumstances, either additional tests are performed in order to fulfill the requirements to use the statistical extrapolation method (Section 3.3.3) or empirically derived assessment factors must be used. It should be recognised that these factors do not have a strong scientific validity and have been rather used as rule of thumb.

Typically, ETVs are calculated from the lowest acute LC_{50} or EC_{50} or, preferably, from the lowest chronic NOEC/L(E)C_x, plus the application of assessment factors that depend on the amount of toxicity data available. It has never been recommended to use fixed “all-purpose” assessment factors. Rather, they should be tempered with whatever information is available in a given situation (Environment Canada 1999). In general, the size of the applied assessment factor will decrease as confidence in the data set increases. The requirements to be fulfilled for the different environmental compartments may differ. Several sets of assessment factors have been

embedded in different regulatory frameworks (EU-TGD 2003; OECD 2004; ECHA 2008; AEE, 2009).

Smaller assessment factors will be used with larger and more relevant data sets (eg, data available for a number of trophic levels, different feeding strategies etc). In most frameworks this is, however, not infinite. If the data set already fulfills the requirements in order to use the lowest assessment factor, extending the data set will not lead to the derivation of a lower assessment factor. In case sufficient data are available, the use of statistical extrapolation methods is preferred.

3.3.3 Calculation of ETV using statistical extrapolation methods (Data-rich substances)

3.3.3.1 Introduction

When a large data set for different taxonomic groups is available, the ETV can be calculated using the statistical extrapolation method in which the susceptibility of a set of species for a given toxicant can be described by some statistical distribution (ie, Species Sensitivity Distribution or SSD). General background information, the derivation of an SSD with regard to the choice of the appropriate distribution model, choice of protection level, and number of data needed are given in Annex A.

Considering that different chemicals have different mechanisms of toxic action (which is the case for metals) and that different organisms react differently to the same chemical, there is a need for flexibility in choosing the appropriate distribution type for developing an SSD. Experts in the field of SSDs recommend against the *a priori* selection of a specific statistical model for an SSD and, rather, suggest that there is considerable latitude for developing appropriate SSDs for various applications (Posthuma et al 2001). However, it is recommended that SSD functions should not be too complex (preferably log-normal distributions are preferred if providing good lower tail fitting and alternatively other models with 2-3 parameters maximally)[11].

In the case of natural elements such as metals, the use of threshold models instead of the commonly used log-normal function can be considered more relevant and scientifically justified. Indeed, assessment of metal SSDs requires consideration of several unique aspects, such as background concentrations, which organisms have evolved with, and essentiality for normal

[11] In statistics, “over fitting” is fitting a statistical model that has too many parameters. An absurd and false model may fit perfectly if the model has enough complexity by comparison to the amount of data available. A perfect fit can therefore always be obtained by using, for example, a high degree polynomial distribution. However, one should not forget that the NOEC values in a SSD represent only a small sample of all sensitivities encountered in an ecosystem and, as such, the true distribution of species sensitivities will always be unknown.

metabolic functions. Metals taken up by active transport have a threshold metal concentration below which the organism cannot uptake the metal from the environment. Accordingly, the *a priori* use of a model such as the normal (or log-normal) distribution, with tails extending to infinity, may result in unrealistically low HC_5 estimates that are within the range of typical background concentrations or, in the case of essential metals, potentially HC_5 estimates that may lie within the range of metal deficiency for some organisms. For this reason, it has been suggested that a threshold model for SSD development may be more appropriate for metals in general, and essential elements in particular (Brix et al 2001; Van Straalen 2002; Van Sprang et al 2004). Van Straalen (2002), for example, found that the triangular distribution provided the best fit of four finite distributions applied to zinc toxicity data, while Brix et al (2001) and Van Sprang et al (2004) used a Pareto model to characterise the threshold response observed in chronic copper and zinc toxicity data, respectively.

Statistical software tools that have been used to calculate SSDs in the framework of risk assessment are the ETX 2.1 program (log-normal distributions) (Van Vlaardingen et al 2004), the Burrlioz program (Burr type III distributions) (Shao 2000) or other more general best fit commercial software packages (@Risk etc) (see Table 6).

Free SSD modeling software	Download link
ETX 2.1	http://www.rivm.nl/rvs/Risicobeoordeling/Modellen_voor_risicobeoordeling/ETX
Burliozz	https://research.csiro.au/software/burrlioz/

Table 6: *Overview of free statistical software*

3.3.3.2 Specific requirements to be fulfilled before using the statistical extrapolation method and practical experiences

In practice, recommendations for the minimum number of species needed to establish a species sensitivity distribution range considerably. Newman et al (2000) determined that between 15 and 55 data points were required for reliable determinations of HC5-50 values. But most meta-analyses that evaluated minimum numbers of species utilise acute toxicity data sets (Newman et al 2000; Duboudin et al 2004a,b; Dowse et al 2013). This is clearly not feasible for regulatory systems that require chronic ecotoxicity data. With regard to the minimum species requirements when using the SSD approach for the aquatic compartment, the London workshop (2001) formulated some recommendations, which are adopted by the European REACH Guidance (European Chemicals Agency 2008). The SSD should contain at least 10 EC_{10}/NOEC values (preferably more than 15) for different species covering at least 8 taxonomic groups (Table 7).

In reality, for some metal/metal compounds and in particular for the sediment and soil compartment, it will be difficult to obtain 10 EC_{10}/NOEC values. In those cases, an SSD could still be constructed as long as the associated sampling uncertainty with the HC_p estimate is properly quantified[12].

Use of SSDs for the water compartment-EU experience

The minimum group requirements to be fulfilled according to the London workshop and the REACH Guidance for the aquatic compartment are summarised in Table 7. Deviations from these recommendations can be made on a case-by-case basis through consideration of sensitive endpoints, sensitive species, mode of toxic action, and/or knowledge from structure-activity considerations.

Taxonomic groups	
1	Fish (usually tested species like salmonids, bluegill, channel catfish etc)
2	A second family in the phylum Chordata (fish, amphibian etc)
3	A crustacean (eg, cladoceran, copepod, ostracod, isopod, amphipod, crayfish etc)
4	*An insect (eg, mayfly, dragonfly, damselfly, stonefly, caddisfly, mosquito, midge etc)*
5	A family in a phylum other than Arthropoda or Chordata (eg, Rotifera, Annelida, Mollusca etc)
6	A family in any order of insect or any phylum not already represented
7	Algae
8	Higher plants
Recommendation: ideally the SSD should cover at least 8 taxonomic groups containing at least 10 NOEC/EC_{10} values (preferably more than 15) for different species (London workshop 2001).	

Table 7: *Minimum taxonomic group requirements for the derivation of ETV for water (freshwater) using the statistical extrapolation technique (London workshop 2001; taken over in ECHA guidance 2008)*

An additional recommendation is that all individual EC_x/NOEC data from one trophic level should not be below the HC_5 estimate. If it appears that all such EC_x/NOEC values are lower than the HC_5, then this could be an indication that a particularly sensitive group exists, implying that some of the underlying assumptions for applying the statistical extrapolation method may not be met. In this respect, bioavailability considerations can also be important.

[12] See Aldenberg and Jaworska (2000) for an example to quantify uncertainty.

Guidance on the minimum number of taxonomic groups needed to apply the statistical extrapolation technique for the terrestrial and sediment compartment is, however, not specifically available because it was not part of the London workshop and is not mentioned in the REACH Guidance documents. For these compartments, the approach has been developed based upon expert judgment and experiences in the framework of the European risk assessments for metals.

Use of SSDs for the soil compartment

The SSD approach has been used for the derivation of an ETV for several metals. Data for a sufficient number of species are generally available to comply with the general recommendations on the minimum number of 10 species to be covered in a SSD (Table 8). Based on this experience, it is recommended that the following taxonomic groups should be covered: at least 2 species of dicotyle plants, belonging to different families, a monocotyle plant, an arthropod, an annelid worm, and microbial processes relating to the carbon and nitrogen cycle.

Taxonomic Group	**Cd**	**Co**	**Cu**	**Mo**	**Ni**	**Pb**	**Zn**
Higher plants (dicotyle)	10 (7 families)	5 (3 families)	5 (3 families)	3 (3 families)	6 (5 families)	3 (3 families)	11 (5 families)
Higher plants (monocotyle)	3 (1 family)	2 (1 family)	4 (1 family)	2 (1 family)	5 (2 families)	7 (1 family)	7 (2 families)
Arthropods	1	1	5 (3 families)	1	2 (1 family)	1	2 (2 families)
Annelida	4 (1 family)	3 (2 families)	4 (2 families)	2 (2 families)	4 (2 families)	3 (1 family)	6 (2 families)
Microbial process – N-cycle	1	1	4	1	2	3	4
Microbial process – C-cycle	2	2	4	2	4	2	5
Other	4	0	2	0	20	2	8
Total	25	14	28	11	43	21	43

Table 8: *Availability of metal toxicity data for direct effects to terrestrial organisms*

Toxicity assays based on microbial functions (eg, respiration or nitrification) differ from the standard single-species tests for plants and invertebrates. Both the tests on single species (plants and invertebrates) and the tests on microbial functions (eg, respiration) can be used to

derive the SSD for the terrestrial compartment. It may be argued that data on microbe-mediated processes and single species tests should be considered separately due to fundamental differences between these tests (functional *vs.* structural test, multi-species *vs.* single species, adapted indigenous microbe community *vs.* laboratory test species, variability of test design and different endpoints, etc). However, the goal of the SSD approach is the protection of the terrestrial ecosystem by protecting its individual constituents. This makes splitting the terrestrial data set redundant. Moreover, protecting microbial functions is considered more relevant compared to protecting individual micro-organisms because the composition of the microbial community can respond strongly to stressors in general (eg, temperature, drought, contamination) in a short time-frame, while potentially still maintaining its functions. In favor of this approach is that each tested community is unique, like each species in the structure-based approach. Thus, a range of such tests yields a range of sensitivities of communities, especially regarding functions that can be taken up as individual points in the SSD. A split of the data set may, however, be required depending on the mode of action making certain species very vulnerable (eg, herbicides/pesticides). If it can be shown that the different types of tests have a similar mode of action for the metal of concern, no split is required and the data from plants/invertebrates/microbial functions should be pooled. For metals, it has been shown that there are no distinct differences in sensitivity between plants, invertebrates, and microbial processes (Kuperman et al 2014). Inclusion of data for additional groups of soil organisms or processes increases the ecological relevance and robustness of the SSD and therefore it is recommended to combine them all into one SSD.

Use of SSDs for the sediment compartment

According to the proceeding of the topical scientific workshop on the principles for environmental risk assessment of the sediment compartment (ECHA 2014), SSD modeling can in theory be considered applicable to obtain effect thresholds for the sediment risk process. The practical applicability of the technique should be carefully evaluated on a case-by-case basis. No general guidelines are available to judge whether a chemical is amenable to SSD modeling or not. According to Vangheluwe et al (2013), applying a SSD for the sediment compartment should take into account (Vangheluwe et al 2013):

1. The expected differences in species richness between sediment and water ecosystems;
2. The different exposure conditions and feeding behaviour of the organisms in the sediment (ingestion of sediment, body wall contact, exposure through pore water, and overlying water);
3. The limitation that very few standardised methods have been established for benthic species overall.

Taxonomic Group	Ni	Cu	Pb	Zn	Cd
Crustaceans (eg, amphipods)	2	2	2	2	/
Insects (eg, diptera)	2	2	2	1	1
Oligochaetes (eg, tubificidae, lumbriculidae)	2	1	2	/	/
Others	1	/	/	/	/
Total	7	5	5	3	1

Table 9: *Availability of metal toxicity data for direct effects to sediment organisms (freshwater sediments)*

For example, for nickel an SSD was established for 7 benthic species representative of different sediment exposure pathways, as well as a variety of feeding strategies and taxonomic groups (Besser et al 2013; Vangheluwe et al 2013). With regard to feeding strategy, a distinction can be made between: (1) surface deposit and filter feeders (eg, *Hyalella azteca*, *chironomids*); (2) burrowing sub-surface feeders (eg, *oligochaetes*); and (3) burrowing species with combined surface and subsurface feeding behavior (eg, *Hexagenia spp.*). According to ECHA (ECHA 2014), the more (ecologically distinct) taxa-- representatives of different feeding strategies and micro-habitats and life forms-- that are contained in the SSD, the more uncertainty is reduced.

Use of SSDs - Experience USA

Compared to the above-mentioned EU-methodology for deriving a PNEC for the freshwater/marine environment (median HC_5 of a SSD based on at least 10 organisms covering 8 different taxonomic groups), the United States Environmental Protection Agency (USEPA) takes a different approach for the derivation of so-called Final Acute and Final Chronic reference Values (FAV, FCV) based on the distribution of the toxicity data. In short, the USEPA calculation method (Stephan et al 1985) takes into account the four data points that have cumulative probabilities closest to 0.05 (if there are less than 59 values, these will always be the four lowest values) and the total number of available data points in the acute or chronic ecotoxicity data set for calculation of the FAV or FCV. The FCV can also be derived by dividing the FAV by a Final Acute-to-Chronic Ratio (FACR). The FCV is then used to determine the Criterion Continuous Concentration (analogous to PNEC) (CCC=FCV/2).

The USEPA uses minimum toxicity database requirements to ensure that the toxicity data collected for a chemical represent a wide taxonomic range of aquatic organisms, which, in turn, is assumed to represent the range of species sensitivities in the natural environment. USEPA

guidance for acute criteria development states that a minimum data set must be available for at least eight different families of aquatic organisms, which is more commonly called the eight-family rule. In the latest version of USEPA guidance, a minimum of eight was chosen to ensure that the four lowest GMAVs (Genus Mean Acute Value) would, by definition, all be in the lowest 50^{th} percentile of available data to limit the amount of extrapolation required to estimate the 5^{th} percentile (FAV).

Similar to the requirements set by ECHA (2008), the minimum data set specified by USEPA for the freshwater environment should represent at least eight families (Table 10). Algal species and higher plants (eg, green alga, duckweed) are not included, and available no-effect levels for these taxonomic groups are discarded when deriving the FCV.

	Taxonomic groups
1	Salmonidae family (Osteichtyes)
2	Second family in Osteichtyes
3	Third family in phylum Chordata
4	Planktonic crustacean
5	Benthic crustacean
6	An aquatic insect
7	Family in phylum other than Chordata
8	Family in any order of insect, or any phylum not already represented

Table 10: *Taxonomic requirements for the derivation of a freshwater Final Acute/Chronic Value (USA)*

Requirements SSD-to specific regions

Given the limited differences between the sensitivity between temperate species and tropical species, there is currently no need to split an SSD between temperate and tropical species. As such, the requirements for populating tropical SSDs will not differ from these for temperate regions but the SSD have to include representatives of the most relevant taxonomic groups present in the region. For example, coral reefs are a typical and important tropical habitat and a tropical marine SSD could include at least one coral species (eg, *Acropora* and *Goniastrea* have been used in ecotoxicity testing).

3.4 Weight-of-Evidence: effects

3.4.1 Introduction

To evaluate the impact on community responses or system functions, results from microcosms, mesocosms, and field studies can provide supporting evidence in a kind of weight-of-evidence approach (WoE) to validate the derived ETV. Toxicity-based extrapolation approaches could in the future be complemented by a suite of ecologically more relevant effect models including population and food web models (De Laender et al 2013). This would help to add the well-needed ecosystem perspective to an environmental risk assessment (De Laender and Janssen 2013)

Weight-of-evidence (WoE) is an often-used phrase 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-upon formalised concept characterised by defined tools and procedures (Weed 2005) and as such, 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 approach can be used in a broad context and is dealt with in a separate fact sheet (Fact Sheet 9). Overall, 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 legislation, the so-called weight-of-evidence 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 weight-of-evidence 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. 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.

Most risk assessment approaches have as their foundation the use of ETVs that 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 risk assessment approach increases, as the relevant database (including local organisms and physico-chemical conditions) decreases. This increasing uncertainty can be countered with additional Lines of Evidence (LoE) that incorporate a field validation component, thereby allowing for the refinement of predictive models (eg, AVS, TOC bioavailability normalisation factors, indigenous species toxicity thresholds, population/community indices). For additional information on utilizing WoE in the field assessment process, see the discussion in the separate WoE Fact Sheet 9.

3.4.2 Using WoE in the ETV derivation process

3.4.2.1 Considering all lines of evidence

In Europe, it is common to derive an ETV to express safe levels (chemical standards) of chemical exposure in water, soil, and sediment. The process for this derivation varies, and can utilise various LoE, such as chronic ecotoxicity data on species, field data, and theoretical models (such as equilibrium partitioning, Biotic Ligand Model). As discussed in the above WoE framework process, the quality of the data, its realism (lack of uncertainty), and the degree to which the ecosystem and relevant receptor responses have been characterised all are important issues to consider in the ETV derivation process. For example, there is substantial uncertainty associated with ETVs 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. Also, using only field-derived empirical guidelines (such as Probable Effect Levels (MacDonald et al 2000) whereby data from a multitude of sites with multiple contaminants are used to predict single chemical threshold effects at other sites, there are unacceptably large uncertainties (over 30%), and require the use of additional LoE (Adams et al 2005). Therefore, to reduce this uncertainty, additional LoE should be used in a sound, relevant, and statistically powerful manner (Wenning et al 2005).

The combination of only two LoE, such as laboratory toxicity data with field-derived effects data, however, greatly reduces uncertainty and adds credibility to the ETV derivation process. This credibility is strengthened with the addition of other useful LoE, such as theoretically sound and proven approaches using organic carbon and acid volatile sulfide (AVS) normalisations of sediment chemistry data for nonpolar organics and divalent cation metals, 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 if conducted properly. The two primary problems that occur in spiking studies are:

1) Lack of adequate equilibration time (ie, partitioning of chemical to sediments) which can greatly increase organism exposure through the water route.

2) A related problem, namely 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 2004). The advantages of the spiking study is that multiple concentrations can be used to derive more accurate threshold levels with associated known variance and that they can be utilised in field settings, thereby allowing for exposures to indigenous populations under natural conditions. Other useful LoE include tissue residue data 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. Accumulation of metal may not equate to adverse effects.

3.4.4.2 Data quality, quantity, and relevance

The quality of the data utilised in the ETV 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 quality assurance/quality control (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, DOC, TSS, 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 2002).

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, PNECs) problematic and inaccurate. However, more data does not necessarily equate to better conclusions, because 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. Thus, 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 have become popular in recent years and have several advantages for determining threshold effect levels (Posthuma et al 2000). 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 not relevant 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 beg the need for additional LoE that are more field-based.

It is apparent from the many issues discussed above that address improving certainty in the determination of threshold effects, ie, ETVs (thus protecting the ecosystem from adverse harm) will require multiple LoE, and should be integrated in a WoE process . **The key LoE of greatest weight 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 would 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 an ETV that has acceptable variance. Often the study may produce unbounded NOEC and 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. Therefore, multiple LoE become essential, combining the field, laboratory, and modeling components to better derive ETVs that are bioavailability based, and using bioavailability normalisation factors and species that are relevant to the ecosystems of concern (see Example 1 below).

Example 1: Use of the Weight of Evidence approach in setting a PNEC freshwater sediment for copper compounds in the EU

In April 2008, the copper industry completed a voluntary risk assessment on copper including the derivation of an ETV, ie, 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 Risk (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 Chemical Agency (ECHA):
(http://echa.europa.eu/chem_data/transit_measures/vrar_en.asp)

Cu is, compared to organic substances, 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 in order to derive a final PNEC. Different lines of evidence (LoE) 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 considering 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.

<u>LoE 1: Benthic toxicity data</u>
Available single species sediment exposure tests resulted in 106 individual high-quality chronic NOEC values for 6 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 insect *Chironomus riparius* (27 individual NOEC values) and the insect *Hexagenia sp.* (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 present and the presence of sulfides. Because both AVS and OC seems 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-$50_{\text{sediment (benthic SSD)}}$ for copper has thus been based on the organic carbon normalised dataset,

using only low AVS sediments and includes 6 species-specific data points (see above) representing 62 NOEC values. The HC5-50 was estimated using the statistical extrapolation methodology. Both best fit distribution as the lognormal distribution were derived.

LoE 2: Use of the Equilibrium Partitioning method

Because metal toxicity in sediments in general is mainly exerted via pore water exposure (eg 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 Kd values obtained from monitoring data, and (b) using the Kd values calculated by the WHAM speciation model.

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 7 benthic species) and the application of following Kd value : the EU median $Kd_{suspended\ solids}$, the EU median $Kd_{sediment}$, 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 organic carbon binding, all values were normalised for their organic carbon content. For each approach, the lowest HC5-50 values (1833 to 3808 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 HC5-50 values. Sediment threshold values and benthic NOEC values are available from the 4 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 organic carbon normalised NOEC (4285 mg Cu/kg OC) is a factor 2.1 (best fit) to 2.5 (log normal) above the $HC5\text{-}50_{\mathbf{sediment\ (benthic\ SSD)}}$. An organic carbon based $HC5\text{-}50_{sediment\ (mesocosm\ SSD)}$ (5^{th} and 95^{th} Confidence limits) was calculated as 3,007 mg Cu/kg OC.

LoE 4: Comparison with terrestrial data

Considering 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 9 microbial endpoints) and soil bioavailability models (pH, OC and CEC normalisations).

Integration of all LoE into the WoE process

The HC5-50 values of all LoE are summarised in Table 11

Line of Evidence	Approach	HC5-50 (organic carbon normalised (mg/Cu/kg OC)	Weight
LoE1: benthic toxicity data			
	SSD best fit & low AVS	2,021	+++
	SSD log-normal & low AVS	1,741	++
LoE2: Equilibrium partitioning			
	EP-WHAM RWC	1,833	+
	EP-KdSS RWC	2,359	+
	EP-KdSED RWC	3,808	+
LoE3: Mesocosms/field data			
	Mesocosm/field SSD lognormal	3,007	++++
LoE4: Terrestrial data			
	Application soil NOEC values and soil bioavailability models	1,723-3,817	±

Table 11: *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 weight-of-evidence approach to assess the remaining uncertainty.

Comparison between the OC normalised HC5-50 values obtained with the EP approach shows that the $HC5\text{-}50_{\text{sediment (EP WHAM)}}$ values are somewhat lower than $HC5\text{-}50_{\text{sediment (EP Sed)}}$ 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 organic carbon sequestration help explain the high variability of those values in the literature. The lowest mesocosm organic carbon normalised NOEC (4,285 mg Cu/kg OC) is a factor 2.1 (best fit) to 2.5 (log normal) above the $HC5\text{-}50_{\textbf{sediment (benthic SSD)}}$. An

organic carbon-based HC5-$50_{sediment\ (mesocosm\ SSD)}$ (5^{th} and 95^{th} Confidence limits) was calculated as 3007 (2204-3743) mg Cu/kg OC. The mesocosm HC5-50 is a factor 1.5 to 1.7 above the derived HC5-$50_{\textbf{sediment (benthic SSD)}}$ (not significant at 0.05 level but significant differences at the 0.01 level). The mesocosm data therefore clearly demonstrate that the HC5-$50_{sediment}$ values, derived through equilibrium partitioning 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 NOEC values/bioavailability models (no significant differences could be detected) between the OC carbon-based HC5-50 values derived. This comparison therefore adds further evidence on the protective nature of the HC5-$50_{sediment\ (benthic\ SSD)}$ values, derived from benthic ecotoxicity tests. In order to evaluate the remaining uncertainty and the need for an additional application factor (AF) on the HC5-50 also the following elements were 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 NIEC 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 Cu/kg dry weight.**

4. REFERENCES

Adams, WJ, Green, AS, Ahlf, W, Brown, SS, Burton, GA, Chadwick, B, Crane, M, Gouguet, R, Ho, KT, Hogstrand, C, Reynoldson, TB, Ringwood, AH, Savitz, JD, Sibley, PK (2005). Using sediment assessment tools and a weight-of-evidence approach. In Wenning, RJ, Batley, GE, Ingersoll, CG, Moore, DW (eds), *Use of Sediment Quality Guidelines and Related Tools for the Assessment of Contaminated Sediments.* SETAC Press, Pensacola, FL., USA, pp 163-226.

AEE (2009). Australian Environment Agency: *Environmental risk assessment guidance manual.*

Agerstrand, M, Küster, A, Bachmann, J, Breitholz, M, Ebert, I, Rechenberg, B, Rudén, C (2011a). Reporting and evaluation criteria as means towards a transparent use of ecotoxicity data for environmental risk assessment of pharmaceuticals. *Environmental Pollution,* 159(10), 2487-2492.

Agerstrand, M, Breitholts, M, Ruden, C (2011b). Comparison of four different methods for reliability evaluation of ecotoxicity data. A case study of non-standard test data used in environmental risk assessments of pharmaceutical substances. *Environmental Sciences* Europe, 23:17.

Aldenberg, T and Slob, W (1993). Confidence limits for hazardous concentrations based on logistically distributed NOEC toxicity data. *Ecotoxicology and Environmental Safety*, 25, 48-63.

Aldenberg, T and Jaworska, JS (2000). Estimation of the hazardous concentration and fraction affected for normally distributed species sensitivity distributions. *Ecotoxicology and Environmental Safety*, 46, 1-18.

Alloway, BJ (ed) (1995). *Metals in soils*. 2nd ed., New York: Blackie Academic & Professional, Chapman and Hall.

ASTM International (2004). Guide for conducting three brood renewal toxicity tests with *Ceriodaphnia dubia*. Guideline E1295 (Annual Book of ASTM Standards 11.05), West Cohockton, PA.

Batley, GE, van Dam, RA, Warne, MSTJ, Chapman, JC, Fox, DR, Hickey, CW Stauber, JL (2014). Deriving Australian and New Zealand Water Quality Guideline Values for Toxicants.

Prepared for the Council of Australian Government's Standing Council on Environment and Water (SCEW).

Besser, JM, Brumbaugh, WG, Ingersoll, CG, Ivey, CD, Kunz, JL, Kemble, NE, Schlekat, CE Garman, ER (2013). Chronic toxicity of nickel-spiked freshwater sediments: Variation in toxicity among eight invertebrate taxa and eight sediments. *Environ. Toxicol. Chem,,* 32(11), 2495-2506.

Brain, P and Cousens, R (1989). An equation to describe dose responses where there is a stimulation of growth at low doses. *Weed Rese*arch, 29, 93-96.

Brix, KV, DeForest, DK, Adams, WJ (2001). Assessing acute and chronic copper risks to freshwater aquatic life using species sensitivity distributions for different taxonomic groups. *Environ. Toxicol. Chem.,* 20(8), 1846-1856.

Buekers, J, Degryse, F, Maes, A, Smolders, E (2008). Modelling the effects of ageing on Cd, Zn, Ni and Cu solubility in soils using an assemblage model. *European Journal of Soil Science, 59*(6), 1160-1170.

Burgess, RM, Berry, WJ, Mount, DR, Di Toro D (2013). Mechanistic sediment quality guidelines based on contaminant bioavailability: equilibrium partitioning sediment benchmarks. *Environ. Toxicol. Chem.*, 24, 2410-2427.

Burton, GA Jr, Chapman, P, Smith, E (2002a). Weight of Evidence Approaches for Assessing Ecosystem Impairment. *Human and Ecological Risk Assessment* 8:1657-1673.

Burton, GA Jr, Batley, GE, Chapman, PM,Forbes, VE, Smith, EP, Reynoldson, T, Schlekat, CE, den Besten, PJ, Bailer, AJ, Green, AS, Dwyer, RL (2002b). A Weight-of-Evidence Framework for Assessing Sediment (Or Other) Contamination: Improving Certainty the Decision-Making Process. *Human and Ecological Risk Assessment,* 8, 1675-1696.

Burton, GA Jr, Greenberg, MS, Rowland, CD, Irvine, CA, Lavoie, DR, Brooker, JA, Eggert, LM, Raymer, DFN, McWilliam, RA (2005). *In situ* exposures using caged organisms: A multi-compartment approach to detect aquatic toxicity and bioaccumulation. *Environ. Pollut*.,134, 133-144.

Brumbaugh, WG, Besser, JM, Ingersoll, CG, Mau, TW, Ivey, CD, Schlekat, CE, Garman, ER (2013). Preparation and characterization of nickel-spiked freshwater sediments for toxicity tests: toward more environmentally realistic nickel partitioning. *Environ. Toxicol. Chem,,* 32 (11), 2482-2492.

Calabrese, EJ (2005). Paradigm lost, paradigm found: the re-emergence of hormesis as a fundamental dose response model in the toxicological sciences. *Environmental Pollution,* 138, 370-412.

CCME (2007). *A protocol for the derivation of Water Quality Guidelines for the Protection of Aquatic Life*. Canadian Council of Ministers of the Environment. Ottawa, ON?

Cedergreen, N, Ritz, C, Streibig, JC (2005). Improved empirical models describing hormesis. *Environ. Toxicol. Chem.*, 24(12), 3166-3172.

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, McDonald, BG, Kickham, PE, McKinnon, S (2006). Note: global geographic differences in marine metals toxicity. *Marine Pollution Bulletin,* 52, 1081-1084.

Checkai, R, Van Genderen, E, Sousa, JP, Stephenson, G, Smolders, E (2014). Deriving site-specific clean-up criteria to protect ecological receptors (plants and soil invertebrates) exposed to metal or metalloid soil contaminants via the direct contact exposure pathway. *Integrated Environmental Assessment and Management,* 10(3), 346-357.

Cullen and Frey (1999). Probabilistic techniques in exposure assessment. A handbook for dealing with variability and uncertainty in Models and Inputs. Plenum Press, New York and London, 335 pages, ISBN 0-306-45956-6.

De Laender, F and Janssen, CR (2013). Brief communication: the ecosystem perspective in ecotoxicology as a way forward for the ecological risk assessment of chemicals. *Integrated Environmental Assessment and Management*, 9(3), e34 (online only),

De Laender, F, Van Sprang, P, Janssen, CR (2013). A re-evaluation of fifteen years of European Risk Assessment using effects models. *Environ. Toxicol. Chem.,* 32(3), 594-601.

OECD (2006). Guidance Document on Simulated Freshwater Lentic Field Tests. OECD Environment Health and Safety Publications N°53.

Dowse, R, Tang, D, Palmer, G, Kefford, B (2013). Risk assessment using the species sensitivity distribution method: data quality versus data quantity. *Environ. Toxicol. Chem.*, 32, 1360-1369.

Duboudin, C, Ciffroy, P, Magaud, H (2004a). Acute-to-chronic species sensitivity distribution extrapolation. *Environ. Toxicol. Chem.*, 23(7), 1774-1785.

Duboudin, C, Ciffroy, P, Magaux, H (2004b). Effects of data manipulation and statistical methods in species sensitivity distributions. *Environ. Toxicol. Chem*., 23. 489-499.

Durrant, C, Stevens, JR, Hogstrand, C, Bury, NR (2011). The effect of metal pollution on the population genetic structure of brown trout (*Salmo trutta* L.) residing in the River Hayle, Cornwall, UK. *Environmental Pollution*, 159, 3595-3603.

Environment Canada (1999). Guidance document on application and interpretation of single species tests in environmental toxicology. EPS 1/RM/34. Ottawa, ON, Canada.

Environment Canada (2005). Guidance document on statistical methods for environmental toxicity tests. EPS 1/RM/46 March 2005. Method development and applications section, Environmental Technology Center, Environment Canada, Ottawa, ON.

European Chemicals Agency (2008). Guidance on information requirements and chemical safety assessment. Chapter R.10: Characterisation of dose [concentration]-response for environment. Available on http://echa.europa.eu/documents/10162/13632/information_requirements_r10_en.pdf.

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

European Chemicals Agency (2011). Guidance on information requirements and chemical safety assessment. Chapter R.4: Evaluation of available information. Available on http://echa.europa.eu/documents/10162/13643/information_requirements_r4_en.pdf

European Chemical Agency (2014). Principles for environmental risk assessment of the sediment compartment: proceeding of the topical scientific workshop. Tarazona, JV, Versonnen, B, Janssen, C, De Laender, F, Vangheluwe, M, Knight, D (eds). Ape, D, Burgess, R, Carere, M, De Knecht, J, Deydier, L, Gawlik, B, Karjalainen, AM, Kukkonen, J, Peijenburg, W, Pellizzato, F, Salvito, D, Schlekat, C, Sibley, P, Simpson, SL, Tyle, H, Waeterschoot, H, Weltje, L (associate eds). ECHA-14-R-13-EN.

Euopean Chemicals Bureau (2001). London workshop (see info on page 53).

European Commission (2011). Technical Guidance for deriving Environmental Quality Standards (TGD-EQS).

EU-TGD (2003). Technical Guidance Document on Risk Assessment in support of Commission Directive 93/67/EEC on Risk Assessment for new notified substances, Commission Regulation (EC) No 1488/94 on Risk Assessment for existing substances and Directive 98/8/EC of the European Parliament and of the Council concerning the placing of biocidal products on the market. Part II: Environment.

EU-Zn Risk Assessment (2006). Risk assessment zinc metal-environmental part. Final draft submitted to the European Chemicals Bureau (Ispra, Italy) by the National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands.

Fairbrother, A and Kapustka, LA (1997). Hazard classification of inorganic substances in terrestrial systems: a discussion paper. ISBN 1-895720-21-4. (JOURNAL NAME?)

Fairbrother, A and McLaughlin, M (2002). Fact Sheet on Environmental Risk Assessment 12: Metalloregions. International Council on Mining and Metals (ICMM), Publ. London, UK. January 2002.

Forbes and Calow (2002). Applying weight of evidence to retrospective ecological risk assessment when quantitative data are limited. Human Ecol. Risk Assess., 8: 1625-1639.

GHS (2003). Globally harmonized system of classification and labeling of chemicals (GHS). United Nations, New York and Geneva, 2003.

Hoffman, GR (2009). A perspective on the scientific, philosophical and policy dimensions of hormesis. *Dose-Response*, 7, 1-51.

Hopkin (1989). *Ecophysiology of metals in terrestrial invertebrates*. London: Elsevier Applied Science.

ISO (1993). International Organisation for Standardisation (ISO). 1993. ISO 11269-1. Soil quality: Determination of the effects of pollutants on soil flora. Part 1: Method for the measurement of inhibition of root growth. International Organization for Standardization, Geneva, Switzerland.

ISO (1995). Soil Quality - Effects of Pollutants on Earthworms (*Eisenia fetida*) – Part 2: Determination of Effects on Reproduction. Draft International Standard, International Organization for Standardization, ISO/DIS 11268-2.2.

ISO (2006). Water Quality- Guidance document on the statistical analysis of ecotoxicity data. **International Organization for Standardization,** ISO/TS20281.

Kapustka, LA, Clements, WH, Ziccardi, L, Paquin, PR, Sprenger, M, Wall D. (2004). Issue paper on the ecological effects of metals. 74pp. USEPA Risk Assessment Forum. Papers addressing scientific issues in the risk assessment of metals.

Kase, R, Agerstrand, M, Moermon, C, Korkaric, M (2014). More transparence and consistency in study evaluations: CRED-criteria for reporting and evaluating ecotoxicity data. Presentation at the multilateral meeting in Dübendorf, Switzerland. June 2014. http://www.oekotoxzentrum.ch/projekte/klimisch/doc/cred_June2014

Kase R. (2015a). CRED-Criteria for Reporting and Evaluating Ecotoxicity data. Part I: Comparison of the Klimisch and CRED methods to evaluate reliability and relevance of ecotoxicity studies. Submitted.

Kase R. (2015b). CRED-Criteria for Reporting and Evaluating Ecotoxicity data. Part II: Practicality and perception of the CRED evaluation method. Submitted.

Kefford, NJ, Zalizniak, L, Warne, MS, Nugegoda, D (2008). Is the integration of hormesis and essentiality into ecotoxicology now opening Pandora's box? *Environmental Pollution,* 151, 516-523.

Klimisch, H, Andreae, M, Tillmann, U (1997). A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regul. Toxicol. Pharm.,* **25**, 1-5.

Kuperman, RG, Siciliano, SD, Römbke, J, Oorts, K (2014). Deriving site-specific soil clean-up values for metals and metalloids: rationale for including protection of soil microbial processes. *Integrated Environmental Assessment and Management*, 10(3), 388-400.

Küster, A, Bachmann, J, Brandt, U, Ebert, I, Hickmann, S, Klein-Goedicke, J, Maack, G, Schmitz, S, Thumm, E, Rechenberg, B (2009). Regulatory demands on data quality for the environmental risk assessment of pharmaceuticals. *Regulatory Toxicology and Pharmacology,* 55, 276-280.

Lee, JS, Lee, BG, Luoma, SN, Yoo, H (2004). Importance of equilibration time in the partitioning and toxicity of zinc in spiked sediment bioassays. *Environ. Toxicol. Chem.*, 23(1), 65-71.

Lock, K, Waegeneers, N, Smolders, E, Criel, P, Van Eeckhout, H, Janssen, CR (2006). Effect of leaching and aging on the bioavailability of lead to the springtail *Folsomia candida*. *Environ. Toxicol. Chem.*, 25, 2006–2010.

Euopean Chemicals Bureau (2001). Report of the expert consultation workshop on statistical extrapolation techniques for environmental effects assessments. London, VK, European Chemicals Bureau, 17-18 January 2001.

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.

Ma,YB, Lombi, E, Nolan, AL, McLaughlin, MJ (2006a). Short-term natural attenuation of copper in soils: Effects of time, temperature, and soil characteristics. *Environ. Toxicol.* Chem.*, 25*(3), 652-658.

Ma, YB, Lombi, E, Oliver, IW, Nolan, AL, McLaughlin, MJ (2006b). Long-term aging of copper added to soils. *Environ. Sci. Tech., 40*(20), 6310-6317.

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

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. *Human Ecol. Risk Assess.,* 2, 277-304.

Mertz, W. 1974. The newer trace elements, chromium, tin, vanadium, nickel and silicon. Proc. Nutr. Soc. **33**: 307–315.

Nannipieri, P, Badalucco, L, Landi, L, Pietramellara, G (1997). Measurement in assessing the risk of chemicals to the ecosystem. In: Zelikoff, JT (ed). *Ecotoxicology: Responses*, *biomarkers and risk assessment,* an OECD workshop. Chapter 32. Fair Haven (NY): SOS Publications. p 507–534.

Newman, MC, Ownby, DR, Mézin, LCA, Powell, DC, Christensen, TRL, Lerberg, SB, Anderson, BA (2000). Applying species-sensitivity distributions in ecological risk assessment: Assumptions of distribution type and sufficient number of species. *Environ. Toxicol. Chem.,* 19(2), 508-515.

OECD (1984). *Guidelines for Testing of Chemicals. No. 207, Earthworm, Acute Toxicity Tests.*

OECD (1992). *OECD Guidelines for Testing of Chemicals No. 210. Fish, Early-life stage toxicity test.*

OECD (1995). *Guidance document for aquatic effects assessment. No. 92.*Environment monograph N°92.

OECD (1998). *OECD Guidelines for Testing of Chemicals. No. 202. Daphnia magna reproduction test.*

OECD (2000). *OECD Guidelines for Testing of Chemicals*. No. 220. *Enchytraeid reproduction test.*

OECD (2001). *OECD Guidelines for the Testing of Chemicals. Proposal for a new guideline 218 sediment-water Chironomid toxicity test using spiked sediment.*

OECD (2002). *Manual for investigation of HPV chemicals.* Section 3.2 Guidance on the developments and use of chemical categories. OECD, Paris.

OECD (2004). *Manual for investigating of HPV chemicals,* September, 2004.

Oorts, K, Bronckaers, H, Smolders, E (2006). Discrepancy of the microbial response to elevated Cu between freshly spiked and long-term contaminated soils. *Environ. Toxicol. Chem.*, 25, 845–853.

Ownby. DR and Newman. MC (2003). Advances in quantitative ion character-activity relationships (QICARs: using metal-ligand binding characteristic to predict metal toxicity). QSAR and Combinatorial Science, 22, 2: 241-246.

Posthuma, L, Traas, TP, De Zwart, D, Suter II, GW (2001). Conceptual and technical outlook on species sensitivity distributions. Pages 475-508 in Posthuma, L, Suter II, GW, Traas, TP (eds). *Species Sensitivity Distributions in Ecotoxicology*. Lewis Publishers, Boca Raton, FL. 587 pp.

Posthuma, L, Suter II, GW, Traas, TP (eds) (2002). *Species Sensitivity Distributions in Ecotoxicology.* Lewis Publishers, Boca Raton, FL.

Reiley MC, Stubblefield WA, Adams WJ, Di Toro DM, Hodson PV, Erickson RJ, Keating FJ Jr. 2003. *Reevaluation of the State of the Science for Water-Quality Criteria Development.* SETAC Press, Pensacola, FL.

Reimann, C and Garret, R (2005). Geochemical background: concept and reality. *Science of the Total Environment,* 350, 12-27.

Salminen, R (Chief-editor); Batista, MJ, Bidovec, M, Demetriades, A, De Vivo, B, De Vos, W, Duris, M, Gilucis, A, Gregorauskiene, V, Halamic, J, Heitzmann, P, Lima, A, Jordan, G, Klaver, G, Klein, P, Lis, J, Locutura, J, Marsina, K, Mazreku, A, O'Connor, PJ, Olsson, SÅ, Ottesen, R-T, Petersell, V, Plant, JA, Reeder, S, Salpeteur, I, Sandström, H, Siewers, U, Steenfelt, A, Tarvainen, T (eds) (2005). *Geochemical Atlas of Europe. Part 1: Background Information, Methodology and Maps.* Espoo, Geological Survey of Finland, 526 pages, 36 figures, 362 maps.

Schabenberger, O, Tharp, BE, Kells, JJ, Penner, D (1999). Statistical tests for hormesis and effective dosages in herbicide dose response. *Agronomy Journal,* 91, 713-721.

Shao, Q (2000). Estimation for hazardous concentrations based on NOEC toxicity data: an alternative approach. *Envirometrics,* 11, 583-595.

SIDS Nickel (2008). SIDS Initial Assessment Profile Nickel. http://webnet.oecd.org/Hpv/UI/handler.axd?id=9ce00cf2-296e-459f-a92e-aec0c9771a7b

Simpson, SL, Angel, BM, Jolley, DF (2004). Metal equilibration and bioavailability in laboratory-contaminated (spiked) sediments used for the development of whole-sediment toxicity tests. *Chemosphere*, 54(5), 597-609.

Stephan, CE, Mount, DI, Hansen, DJ, Gentile, JH, Chapman, GA Brungs, WA (1985). *Guidelines fo*r *Deriving Numerical National Water Quality Criteria for the Protection of Aquatic Organisms and Their Uses.* NTIS No. PB85-227049. EPA 440-5-85-501. EPA Environmental Research Laboratory, Duluth, MN.

Suter II, GW (2002). North American history of species sensitivity distributions. Pages 11-17 *in* Posthuma, L, Suter II, GW Traas, TP (eds). *Species Sensitivity Distributions in Ecotoxicology*. Lewis Publishers, London, UK.

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.

Uren, WT, Bury, N, van Aerle, R, Santos, E (2013). Global transcriptome profiling reveals molecular mechanisms of metal tolerance in a chronically exposed wild population of brown trout. *Environ. Sci. Tech.*, *47*, 8869–8877.

USEPA (1999a). Environmental Protection Agency. 1999 Update of Ambient Water Quality Criteria for Ammonia, EPA-822-R99-014. US Environmental Protection Agency, Washington DC.

USEPA (1999b). Draft Risk Assessment Guidance for Super Fund. Volume 3-Part A, process for conducting probabilistic risk assessment. US Environmental Protection Agency, Washington DC.

USEPA (2004). Status and Future Directions of the High Production Volume Challenge Programme. US Environmental Protection Agency, Washington DC.

USEPA (2009). Draft 2009 update Aquatic Life Ambient Water Quality Criteria for Ammonia-freshwater. EPA-822-D-09-001. US Environmental Protection Agency, Washington DC.

Van Assche, F, van Tilborg, W Waeterschoot, H (1996). Environmental risk assessment for essential elements, Case study Zinc. In "*Report of the International Workshop on Risk Assessment of Metals and their Inorganic Compounds*" ICME, Ottawa, Canada. Publ. P/ 171-180.

Van Assche, F, Van Tilborg, W, Waeterschoot, H (1997). Environmental risk assessment for essential elements. Case study: Zinc. International Zinc Association, Brussels, Belgium.

Van Gestel, CAM, McGrath, SP, Smolders, E, Ortiz, MD, Borgman, E, Verweij, RA, Buekers, J, Oorts, K (2012). Effect of long-term equilibration on the toxicity of molybdenum to soil organisms. *Environmental Pollution*, 162, 1-7.

Van Leeuwen, K., Emans, HJ, Van de Plassche, E, Canton, H (1994). Chapter 27: The role of field tests in hazard assessment of chemicals. P. 425-437. In: *Freshwater field tests for hazard assessment of -chemicals.* Hill, I, Heimbach, F, Leeuwangh, P, Matthiessen, P (eds). Lewis Publishers, Boca Raton, FL.

Van Ewijk, PH and Hoekstra, JA (1993). Calculation of the EC50 and its confidence interval. *Ecotoxicology and Environmental Safety,* 25, 25-32.

Vangheluwe, M, Verdonck, F, Besser, J, Brumbaugh, W, Ingersoll, C, Schlekat, C, Garman, E (2013). Improving sediment quality guidelines for nickel: development and application of predictive bioavailability models to assess chronic toxicity of nickel in freshwater sediments. *Environ. Toxicol. Chem.,* 32(11), 2507-2519.

Van Sprang, PA, Verdonck, FAM, Vanrolleghem, PA, Vangheluwe, ML, Janssen, CR (2004). Probabilistic environmental risk assessment of zinc in Dutch surface waters. *Environ. Toxicol. Chem*., 23(12), 2993-3002.

Van Straalen, NM (2002). Threshold models for species sensitivity distributions applied to aquatic risk assessment for zinc. *Environ. Toxicol. Pharmacol.,* 11, 167-172.

Van Straalen, NM and van Leeuwen, CJ (2002). European history of species sensitivity distributions. Pages 19-34 *in* Posthuma, L, Suter II, GW, Traas, TP (eds). *Species Sensitivity Distributions in Ecotoxicology*. Lewis Publishers, London, UK.

Van Vlaardingen, PLA, Traas, TP, Wintersen, AM, Aldenberg, T (2004). ETX 2.0. A program to calculate hazardous concentrations and fraction affected, based on normally distributed toxicity data, Bilthoven, The Netherlands: National Institute for Public Health and the Environment (RIVM). Report N° 601501028/2004, 68pp.

Verdonck, FAM, Souren A, Van Asselt MBA, ML, Van Sprang, PA and Vanrolleghem PA (2006). Improving uncertainty analysis in European Union Risk Assessment of Chemicals. Integrated Environmental Assessment and Management, 3 (3): 333-343.

Vom Saal, F and Myers, JP (2010). Good laboratory practices are not synonymous with good scientific practices, accurate reporting or valid data. *Environmental Health Perspectives Correspondence*, 118(2), A60.

Walker, JD, Enache, M, Dearden, JC (2003). Quantitative Cationic-Activity relationships for predicting toxicity of metals. *Environ. Toxicol. Chem.,* 22(8), 1916-1935.

Wang, F, Goulet, RR, Chapman, PM (2004). Testing sediment biological effects with the freshwater amphipod *Hyalella azteca*: The gap between laboratory and nature. *Chemosphere,* 57(11), 1713-24.

Wang, X, Yan, Z, Liu, Z, Zhang, C, Wang, W Li, H (2014a). Comparison of species distributions for species from China and the USA. *Environ. Sci. Pollut. Res.,* 21, 168-176.

Wang, Z, Kwok, KW, Lui, GC, Zhou, GJ, Lee, JS, Lam, MH, Leung, KM (2014b). The difference between temperate and tropical saltwater species' acute sensitivity to chemicals is relatively small. *Chemosphere*, 105, 31-43.

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

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

Woodward, DF, Hansen, JA, Bergman, HL, Delonay, AJ Little, EE (1995). Brown trout avoidance of metals in water characteristic of the Clark Fork River, Montana. *Canadian Journal of Fisheries and Aquatic Science,* 52, 2031-2037.

Yungsong, M, Fengchang, W, Cheng, C, Yuedan, L, Xiaoli, Z Haiqing, L (2014). Predicting criteria continuous concentrations of 34 metals or metalloids by use of quantitative ion character activity relationships-species sensitivity distributions (QICAR-SSD) model. *Environmental Pollution,* 188, 50-55.

Zajdlik, BA (2006). Potential Statistical Models for Describing Species Sensitivity Distributions. Prepared for the Canadian Council of Ministers of the Environment, Ottawa, Canada. January 2006.

ANNEX A: USE OF SPECIES SENSITIVITY DISTRIBUTIONS

A1. Introduction

A SSD can be visualized as a cumulative distribution function (Figure A1). The cumulative distribution function curve follows the distribution of the sensitivity data obtained from ecotoxicological testing, plotting effect concentrations derived from acute or chronic toxicity tests, for example LC_{50} and EC_{10} values, respectively.

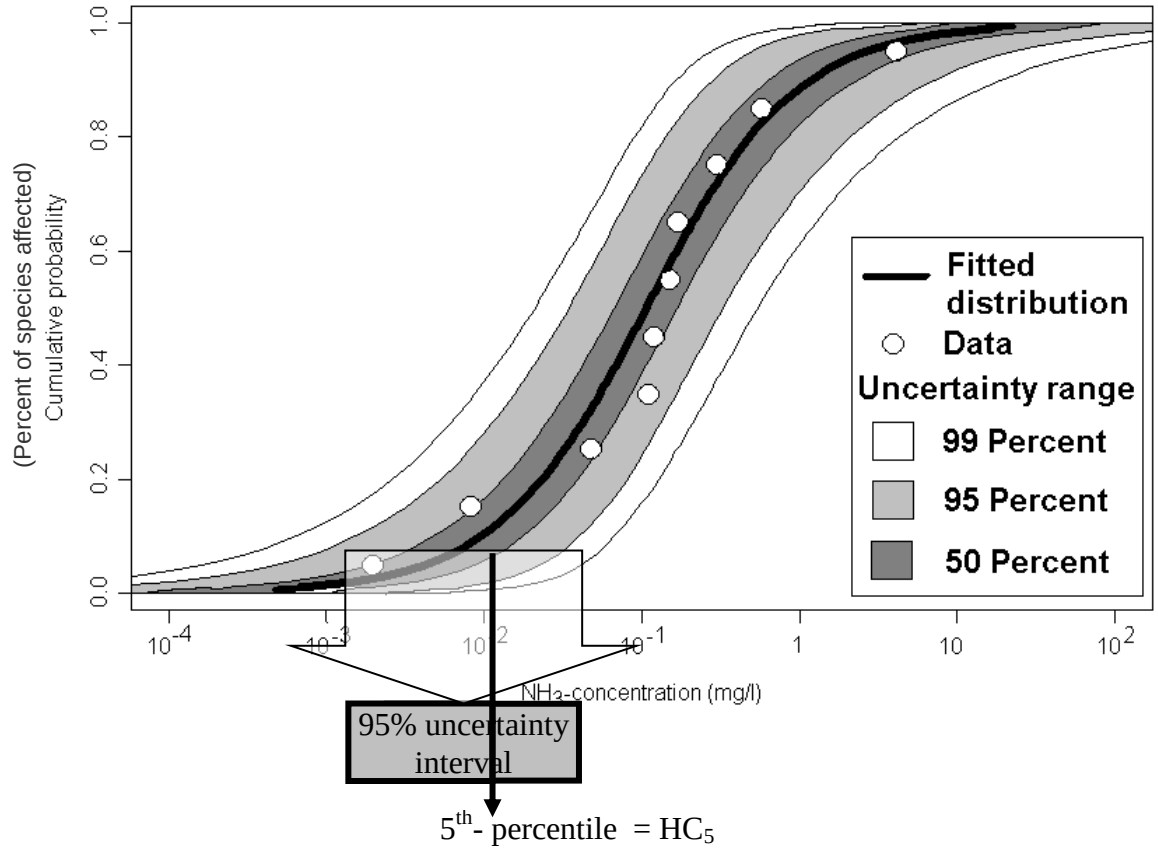

Figure A1: *Example of a SSD (Species Sensitivity Distribution – log-logistic distribution) with uncertainty band and its HC_5 (Hazardous Concentration at 5%)*

SSDs were originally proposed to derive environmental quality standards in the late 1970s and mid-1980s in the United States and Europe, respectively, and their importance in ecotoxicity evaluations has steadily grown (Van Straalen and van Leeuwen 2002; Suter II 2002). A more recent application is the use of SSDs in environmental risk assessments (ERA) The most common current approach is to derive the ETV (eg, Predicted No Effect Concentration, PNEC) from the 5th percentile of SSD (EU-TGD 1995) as shown in Figure A1. Historically that value is known as Hazardous Concentration at p-protection level or HC_p. A cut-off percentage p is

chosen (to protect 1-p percent of species), and the desired "safe" concentration HC_p is calculated. The 5th percentile of a chronic toxicity distribution has been chosen in the earliest methods as a concentration that is protective for most species in a community (namely 1-p %), but the value of p is a policy decision, not science. In popular use of the method, the complementary value of p has become known as the 95% (100-p) protection criterion. Researchers also started to determine a confidence or uncertainty interval on the HC_5 (Figure A1). This was mainly done because the median HC_5 is a conservative estimate of the HC_5 calculated without uncertainty (Aldenberg and Slob, 1993). Note that Aldenberg and Jaworska (2000) extended the calculation of uncertainty to both HC_p and p at a given concentration. A confidence or uncertainty interval can quantify the sampling error in the HC_5 estimate.

Intensive discussions have taken place on principles, statistics, assumptions, data limitations, and the use of SSDs in the field of risk assessments (Posthuma et al 2002; Forbes and Calow 2002). Because an SSD is populated by a variety of chronic endpoints, it has been questioned if using solely an SSD for establishing an acceptable risk level is appropriate in a risk assessment context where the main aim is protecting ecosystem structure and function. The ambiguity in the relationship of SSDs to chronic assessment endpoints is due in part to the lack of guidance from regulatory agencies because the conventional chronic endpoints represent rather thresholds for statistical significance and generally lack biological interpretation. These concerns refer, however, to the use of chronic endpoints in general and pertain also to other approaches such as the assessment factor approach.

Most critiques have tended to focus on the technical aspects of constructing an SSD avoiding the inferential difficulties of defining and estimating assessment endpoints. For example, the choice of distribution functions, dependence of SSDs on the amount and quality of available data, and the choice of protection level have been particularly debated.

Despite the criticisms, the SSD concept is currently used in an array of decision-making processes. Improvements in SSD design and uses can be traced to three important topics (Posthuma et al, 2002). First, the specificity of SSDs can be improved by tailoring the fundamental design of the analysis to the problem using ecological, toxicological, and environmental chemical information. This includes ecological aspects regarding test endpoint, identifying and correcting for nonrandom species selection, correcting for lack of independence between toxicity data, ecological aspects regarding auto-ecological features of species, and biogeography and statistical aspects. Second, SSD techniques and associated techniques that are applied in the assessment can yield improved assessment accuracy, by using concepts from environmental chemistry, ecology, toxicology, biogeography, and taxonomy. Third, the

fundamental statistical techniques can be improved, and regularities in toxicity databases can be used to address the problems of small sample size. The evolution of SSDs along these lines may limit the relevance of criticisms for certain SSD approaches.

A2 Choice of the appropriate distribution model

Numerous methods have been proposed for developing species sensitivity distributions (SSDs) and there is no consensus on the most appropriate method. One of the key aspects is the selection of an appropriate distribution model. Many users of SSDs simply employ a standard distribution such as the log-logistic and the log-normal distribution because these have been historically used. In selecting these functions, statistical arguments have been used more frequently than ecological arguments. Aldenberg and Slob (1993) chose the logistic function based on its inherent properties rather than their fit to data. The model is more conservative than the normal distribution (generates lower HC5 values) and is computationally tractable (Posthuma et al 2002).

Newman et al (2000) evaluated 30 published toxicity data sets and found that the null hypothesis of a log-normal distribution was rejected (α = 0.05) in one-half of the data sets according to the Shapiro-Wilk's test. The authors cautioned that the defensibility of the SSD approach may be compromised if a fundamental assumption behind the approach is frequently violated (ie, assumption that the data are log-normally distributed). In cases where the data are not log-normally distributed, the use of an SSD based on a log-normal approach would not be defensible and can lead to SSDs that badly fit the data and hence cast doubt on the appropriateness of the method.

It is preferable to select functions based on goodness-of-fit tests or other statistical comparisons of alternative functions. According to Zajdlik (2006), the Canadian Council of Ministers of the Environment (CCME) requires using formal tests for goodness of fit in conjunction with graphical assessments of goodness of fit. Goodness-of-fit tests (eg, Anderson-Darling (A-D) and Kolgomorov-Smirnov tests (K-S)) are formal statistical tests of the hypothesis that the data represent an independent sample from an assumed distribution. These tests involve a comparison between the actual data and the theoretical distribution under consideration. The calculated goodness-of-fit statistic measures how good the fit is: critical values are calculated and used in order to determine whether a fitted distribution should be accepted or rejected at a specific level of confidence. Typically, these values depend on the type of distribution fit, the number of data points, and the confidence interval. The level at which one distinguishes between likely and unlikely values of the test statistic is a matter of judgment. A significance

level of 0.05 is most often used, implying that a value of the test statistic below the 95th percentile of the distribution for the statistic is acceptable and leads to the inability to reject the hypothesis. A value of the calculated A-D/K-S statistic above the 95th percentile of the distribution leads to the rejection of the null hypothesis, ie, the distribution is not a good fit (Cullen and Frey 1999). In case of lack of fit at the 95% confidence level, the statistical extrapolation method should not be used.

The (A-D) test places most emphasis on tail values whereas the (K-S) test investigates the data fit for the whole distribution curve to the same extent. Care must be taken when evaluating results of best-fit analyses, because one goodness of fit test statistic (eg, A-D) may indicate that one distribution offers the best fit, while another goodness of fit test statistic (e.g, K-S) may indicate that a different distribution has the best fit. This can influence the choice of the distribution, and also the derivation of the HC_5.

For the purpose of deriving the HC_5 estimate (ie hazardous concentration at which above 95% of the species is protected) to be used in a risk assessment context, preference could be given to the outcome of the A-D test because it places more emphasis on tail values and as such reduces uncertainty in this estimate. If the whole SSD is used as in the case of estimating the "Potentially Affected Fraction", it could be argued from an ecological viewpoint that the K-S test statistic is equally important. Anyway, the left tail of the distribution should always be analysed carefully. If a subgroup of species can be identified as particularly sensitive, the role should be assessed of this species in terms of their function in the ecosystem.

A3 Choice of the protection level (HC_p)

The question arises about which HC_p level of protection is most appropriate. Several choices can be made to derive toxicity thresholds from SSDs (ie, selection and combinations of EC_x and HC_p for example a HC5 from a distribution consisting of EC50 values or EC20 values). There is no proof yet regarding which of these combinations most likely indicates incipient toxicity in the field, ie, is a higher number of species affected reasonable when the adverse effect per species is relatively small, or is it better to have a larger effect on fewer species? For example, the HC_5 of EC_{50} values may be an index of substantial stress (EC_{50}) on a few number of species (5th percentile), whereas the HC_{50} of EC_{10} values is an index of relatively weak stress (EC_{10}) on a larger number (50th percentile) of species. Generally, the choice of the HC_p seems to have a stronger effect on the outcome than the selection of the EC_x (Checkai et al 2014). The selection of both EC_x and HC_p values is not a scientifically based decision but is a regulatory choice and determines the level of effect on species populations that is acceptable. It should be noted that

for small p, HC_p values are often a statistical extrapolation, not interpolation of the data, and that these become increasingly dependent on the assumed curve fitted to the SSD, especially when the number of species is small.

A4 Number of data

The number of data to construct SSDs may vary widely, between a few data (n > 3) to more than 50 or 100 sensitivity values (for data-rich metals). An appropriate question to consider while evaluating the data set as a candidate for the statistical extrapolation approach is 'how many data are needed?' to fit a Species Sensitivity Distribution (SSD) model with sufficient confidence using all available acute/chronic NOEC values as input.

Generally, the larger the sample size, the greater one's confidence in the choice of a probability distribution and the corresponding estimates. Conversely, for small sample sizes, goodness-of-fit statistics will often fail to reject any of the hypothesised probability distribution function. In general, there is no rule of thumb for the minimum sample size needed to specify a SSD. Increasing sample size may, however be an important consideration when making decisions about uncertainty (US EPA 1999). Nonparametric approaches for estimating 5^{th} percentiles can be unreliable at small sample sizes (N < 10) because assumptions on extrapolating below the smallest data point are difficult to make.

A sufficient number of data are needed as a prerequisite for selecting a distribution function with adequate confidence but this is not the only requirement that has to be fulfilled in order to use the SSD approach properly. First, under the intrinsic assumption that the available ecotoxicological data are an independent and identically distributed sample of the real SSD, the available data should be representative in terms of ecological relevance (eg, include key species), and include the appropriate number of taxonomic groups and trophic levels. Secondly, the more entry points (number of NOEC values or $L(E)C_{50}$ values), the more precise the estimation of the HC_5 will be. Furthermore, data quality is equally important as data quantity. Adding more species to a SSD will have no or little impact on the spread of the SSD if the SSD is already based on a representative sample of species. However, by adding more species, the uncertainty on the HC_p value will decrease (Figure A2); The 90% uncertainty bound decreases with increasing number of data points for a hypothetical but realistic situation.

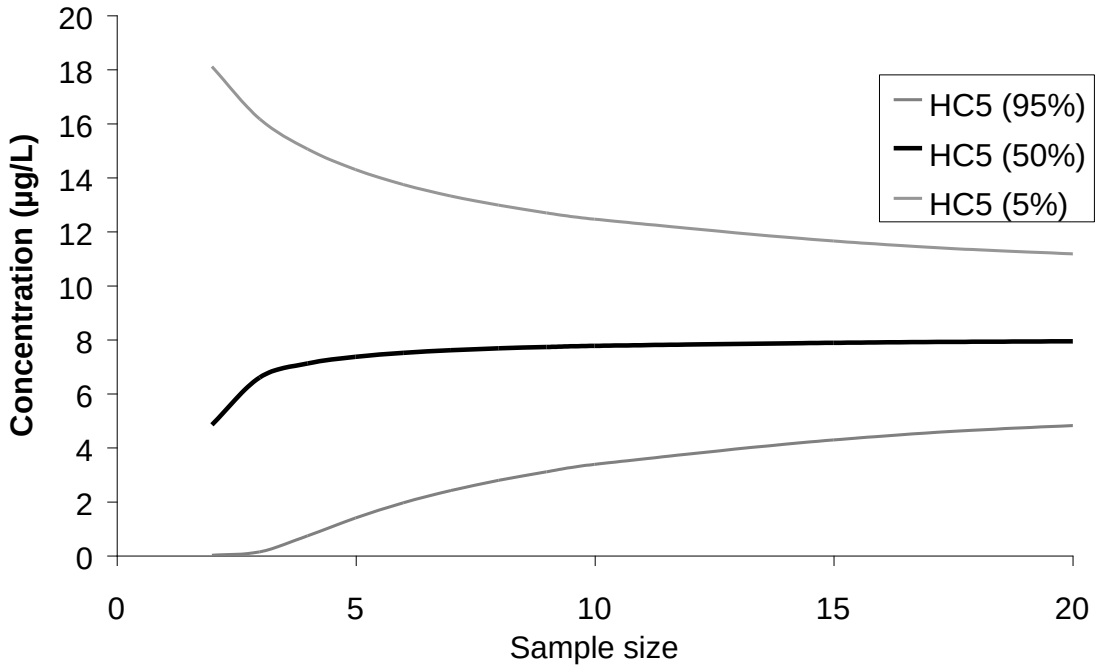

Figure A2: *Relationship between the 90% uncertainty band of the HC_5 estimate and the sample size; where hypothetical case (Lognormal distribution, mean = 1.4, SD = 0.3)*

A5 Data treatment

Where multiple reliable data are available for one species/endpoint and it is apparent from the data that the observed difference in test results for one species is due to differences in bioavailability, it is recommended to process (eg, geometric mean, normalisation) prior to the SSD fitting. If no bioavailability tools have yet been developed for the metal/metal compound under consideration, a pre-selection should be performed in relation to realistic environmental parameters for the compartment under investigation.

As such, the potential bias introduced through the over-representation of ecotoxicological data from one particular species, included in the SSD without further processing, can be avoided[13]. For a risk assessment, the focus should be on interspecies variability and not on intraspecies variability.

Another approach, that can be applied in order to limit the impact of only a few species on the outcome of the entire SSD, is a weighted analysis that takes into account redundant data for each species, so that all data are used, and intra-species variation is taken into account, but no species is given more importance than another, and all data points contribute evenly to the

[13] Generally, internationally acknowledged test species such as *Daphnia magna*, *Pimephales promelas*, or *Raphidocelis subcapitata* generate a large amount of toxicity data, and can thus have much greater importance in the SSDs if all data points are included without further processing

SSD. It is a more complex approach and so far there is, however, no or little experience with this approach under a regulatory framework.

A6 Uncertainty management

Once a HC_p is chosen, it is recommended to look at the remaining uncertainty. According to the REACH Guidance (ECHA 2008), an additional assessment factor on the HC5 value could be applied if deemed appropriate. Typically, in defining assessment factors (AF), the size of an assessment factor depends on the confidence with which an ETV can be derived from the available data. This confidence increases if data are available on the toxicity to organisms at a number of trophic levels, taxonomic groups, and with lifestyles representing various feeding strategies. Thus, lower assessment factors can be used with larger and more relevant data sets than a base-set data." Typically, the following points are considered in the uncertainty assessment of a SSD under the REACH framework:

- The overall quality of the database and the end-points covered, eg, if all the data are generated from "true" chronic studies;
- The range in abiotic factors of exposure media covered;
- The diversity and representativeness of the taxonomic groups covered by the database;
- Statistical uncertainties around the HC_5, eg, reflected in the goodness-of-fit or the size of confidence interval around the 5^{th} percentile;
- Evaluation of EC_{10}/NOEC values below the $HC_{5;}$
- Comparisons between field/mesocosm studies and the 5^{th} percentile to evaluate the laboratory to field extrapolation (see also Section 3. 4 Weight of Evidence).

Thus, among other factors, the size of the assessment factor is mainly driven by the number of species (covering sufficient taxonomic groups). A way forward to objectively evaluate and if needed define an assessment factor to be applied on an HC_5 is looking at the decrease in confidence interval surrounding the HC_5 in function of the number of entries in an SSD. In this regard, it has been suggested by Verdonck et al (2007) to use the ratio between the HC_5 (50 %) and HC_5 (5%) as a surrogate for the assessment factor to cover uncertainty due to limited number of species (assuming that several taxonomic groups are covered).

During this process, care should be taken that the level of conservatism embedded in the recommended assessment factors is consistent with the relation to the number of species used, which can be done by recalibrating the factor according to the widely-accepted level of

conservatism already in use (eg in the EU a factor of 10 for 3 chronic EC_{10} or NOEC data points is used in the classical AF approach).