Environ Sci Pollut Res (2014) 21:33–50
DOI 10.1007/s11356-013-2204-x

Issues to consider in the derivation of water quality benchmarks for the protection of aquatic life

Uwe Schneider

Received: 15 February 2013 / Accepted: 30 September 2013 / Published online: 15 October 2013

Abstract While water quality benchmarks for the protection of aquatic life have been in use in some jurisdictions for several decades (USA, Canada, several European countries), more and more countries are now setting up their own national water quality benchmark development programs. In doing so, they either adopt an existing method from another jurisdiction, update on an existing approach, or develop their own new derivation method. Each approach has its own advantages and disadvantages, and many issues have to be addressed when setting up a water quality benchmark development program or when deriving a water quality benchmark. Each of these tasks requires a special expertise. They may seem simple, but are complex in their details. The intention of this paper was to provide some guidance for this process of water quality benchmark development on the program level, for the derivation methodology development, and in the actual benchmark derivation step, as well as to point out some issues (notably the inclusion of adapted populations and cryptic species and points to consider in the use of the species sensitivity distribution approach) and future opportunities (an international data repository and international collaboration in water quality benchmark development).

Keywords Water quality benchmark · Water quality standard · Guideline · Criteria · Development protocol · Species sensitivity distribution · SSD · Toxicity test evaluation · Cryptic species · Adapted population · Genetic fingerprinting · International data repository

Responsible editor: Philippe Garrigues

U. Schneider (✉)
1944 Forced Road, Ottawa, Ontario, Canada K0A 3H0
e-mail: hots@295.ca

Introduction

The development of water quality benchmarks for the protection of aquatic life has a long history in some North American and European jurisdictions. Not every jurisdiction followed the same approach; different expertise has been developed and a lot of experience has been gained in these countries. Currently, more and more countries are setting up their own national water quality benchmark development programs, and to do this, they have three options available: adopt an existing method from another jurisdiction, update on an existing approach, or develop their own new derivation method. While adopting an existing method is quick and easy, it does not allow for adaptation to the particular jurisdictional needs and the incorporation of newer scientific understanding. Updating of an existing method allows incorporation of new science and adaptation to the jurisdictional requirements, but also requires more resources, time, and an understanding of the relevant issues. If the updating is done thoroughly, it will easily transition into the third option. Developing their own derivation method gives complete freedom to incorporate new science and adapt to particular jurisdictional needs, but is also the most time-consuming and labor-intensive option. Alternatively, jurisdictions may opt to not develop their own derivation method but rather adopt already published water quality benchmarks. While this is again quick and easy, there is the risk that the adopted benchmark value is not suitable or appropriate for the particular jurisdictional needs and is ecologically not relevant. Also, without detailed analysis, there is the danger of adopting a value with errors or low scientific defensibility. While each approach has its own advantages and disadvantages, many issues have to be addressed when setting up a water quality benchmark development program or adopting published values. The intention of this paper was to provide some guidance for this process not only on the program level but also for the methodology development, and in the actual

benchmark derivation step, based on the experiences gained from the recent development of the new Canadian Water Quality Guidelines derivation protocol (CCME 2007a).

The water quality benchmark development triad

Generally, a water quality benchmark program can be separated into three distinct categories: the Authoring Organization, the Derivation Protocol, and the actual Water Quality Benchmarks. Each of these has its purpose and associated challenges and will be examined in detail here.

The authoring organization

The authoring organization is the managing body and is in charge of the broad, overall development program of the water quality benchmarks. It establishes the program; sets the mandate; and looks after the continuation, the funding, the peer review and approval process, the distribution and publication of the benchmarks, and, where applicable, their implementation and execution. In the beginning, it sets in place the designers of the water quality benchmark development program and assembles the necessary expertise. It must be understood that the development of scientifically defensible water quality benchmarks is a challenging and time-consuming task, requiring expertise in many areas beyond environmental chemistry, aquatic toxicology, aquatic ecology, and statistics.

Due to the broad nature of the various tasks, they may be spread out and delegated over two or more management levels and/or horizontal layers in the authoring organization. This has happened, for example, in Canada where the Canadian Council of Ministers of the Environment (CCME) created a high-level board (the Environmental Planning and Protection Committee, made up of representatives of the provincial, territorial, and federal Ministries of the Environment) is in charge of the broader aspects of mandate and funding; a mid-level committee (the Water Quality Task Group, also consisting of representatives of the various provincial, territorial, and the federal Ministries of the Environment) for the mid-term planning, peer review, approval process, and the publication; and a technical group [the National Guidelines and Standards Office, within the federal Department of the Environment (Environment Canada)] for the short-term planning, technical expertise, and actual benchmark derivation (CCME 2013a).

Equally broad is the management structure, for example, in Australia and New Zealand. Here, the Australian National Water Quality Management Strategy (NWQMS) is a joint strategy originally developed by two Ministerial Councils—the former Agriculture and Resources Management Council of Australia and New Zealand (ARMCANZ) and the former Australian and New Zealand Environment and Conservation Council (ANZECC). Since 1992, the NWQMS has been developed by the Australian and New Zealand Governments in cooperation with state and territory governments. Ongoing development is currently overseen by the Standing Council on Environment and Water and the National Health and Medical Research Council. Under their strategy, the ANZECC Standing Committee on Environmental Protection develops and publishes the Australian Water Quality Guidelines for Fresh and Marine Waters (Australian Government 2013).

Decisions on program setup

The designers of such a water quality benchmark program have the choice to either tailor their program after another jurisdiction or design their own new development program. However, when copying or adapting an existing program, they should have a good understanding of the origin of this program. They also have to consider many fundamental issues pertaining to their own country or countries. Aspects like the jurisdictional status [is the program for a single-jurisdictional country, a multi-jurisdictional country (e.g., a federation), or a multi-country union?]; the legal status (are the resulting benchmarks legally binding thresholds or guidance values?); funding security (for the program as well as for the derivation of individual benchmarks); the geographical span (arctic, temperate, and/or tropical); application area (freshwater—lakes, rivers; estuarine; marine—coastline, open seas); application site (end-of-pipe locale, within or outside the mixing zone, or ambient environment); and environmental status (i.e., are the water quality benchmarks applied to highly developed locales or to pristine areas?) have to be clearly defined as they greatly influence not only the scope of the new program but also many technical and scientific details of the actual derivation method.

The authoring organization also has to define the scope of the water quality benchmarks. The scope can be limited to protect aquatic organisms or expanded to include consumers of aquatic biota (wildlife and humans). It has to clarify the goal and purpose of the benchmarks [e.g., protection of aquatic life for the sake of the environment (i.e., broad, e.g., Canada: CCME 2007a) or protection of some aquatic species for human consumption (i.e., narrow, e.g., Japan: Yamazaki 2011)] and the desired protection level (protect all aquatic species or only some species; protect individuals, species, or ecosystems; protect all the time or only some of the time). It has to decide on allowing exceedances (when, how often) and define what is an exceedance (size, frequency, magnitude, spatial extent). It has to set clear definitions for these aforementioned terms to aid in their understanding and interpretation. Failure to do this during the beginning will lead to large problems and difficulties later on in the process: during the drafting of the derivation protocol, the derivation of the water quality benchmarks, and finally in their application and use.

For example, the stated goal may be "to protect aquatic life from harm"—it is now essential to clearly define what is meant by the terms "protect," "harm," and "aquatic life." A scientist, lawyer, or layperson likely will have a vastly different understanding and interpretation of these terms. Different dictionaries and different jurisdictions provide differing definitions for these simple terms. Equally, the level of protection must be clearly identified to avoid confusion and endless debates later on. Is the benchmark intended to protect at the ecosystem function level, the population level, the species level, or even at the individual level? Clarification on these points will also greatly aid in the decision on how to actually derive a water quality benchmark as the "intended scope of the benchmark" dictates many processes and limitations on the scientific method (i.e., the derivation protocol) used to calculate the benchmark value.

Equally important is the decision to keep the water quality benchmarks as purely science-derived values or to incorporate technological or socioeconomic aspects in their derivation. For example, in Canada, the Water Quality Guidelines for the Protection of Aquatic Life (i.e., applied to ambient waters) are purely science-derived values, while the Guidelines for Canadian Drinking Water Quality combine scientific, technological, and socioeconomic aspects. Ideally, water quality benchmarks are science-based (i.e., derived using aquatic toxicity data); however, in the absence of toxicity information, interim benchmarks may also be derived using environmental levels (usually obtained from environmental monitoring programs). Sometimes science-based values are modified due to technological limits [e.g., toxic effects of a contaminant are at or even below the analytical detection limit and a science-based benchmark would, therefore, be below the detection limit, or a benchmark cannot be adhered to with available pollution abatement techniques (this may apply more to effluent limits or drinking water quality guidelines than ambient water quality benchmarks)]. Furthermore, consideration of socioeconomic aspects, such as the cost of achieving a benchmark in a certain area, or localized societal decisions, such as deriving benchmarks for already degraded ecosystems, will lead to modifications of existing benchmarks or derivation of situation-specific thresholds. In some jurisdictions (e.g., Canada; CCME 2003), this realization has led to the creation of two (or more) parallel benchmark development programs, one aimed for example at deriving generic, science-based, jurisdiction-wide (i.e., "national") water quality benchmarks and the other designed to derive situation- and/or site-specific water quality benchmarks (incorporating toxicological, site-specific ecological, technological, and/or socioeconomic aspects in combination or separately). Furthermore, as already mentioned, water quality benchmarks can have different legal standings; they can either be voluntary guidance values or legally enforceable thresholds. This, too, will influence the overall development process and the particular derivation method for a benchmark. It may be prudent for a jurisdiction to keep the derivation, implementation, and remediation measures in separate processes.

Terminology

With respect to water quality benchmarks, unfortunately, a different terminology is being used across jurisdictions around the world, often with conflicting definitions and interpretations. Terminology may even differ within the same country. Terms like guideline, criterion, standard, objective, limit, threshold, trigger value, and benchmark are all used inconsistently, and sometimes interchangeably. What is a "guideline" in one jurisdiction is a "criterion" in another, and a "standard" in a third. However, in many but not all jurisdictions, the term "water quality guideline" [e.g., Canada (national), Autralia, New Zealand, etc.] or "water quality criterion" (e.g., USA) is given to a voluntary guidance value; "water quality standard" is used for a legally enforceable benchmark (e.g., Canada, USA), while the term "water quality objective" is applied when technological or socioeconomic aspects are incorporated [e.g., Canada (national)]. However, there are many exceptions; for example, within Canada, the Province of Ontario had published its "Provincial Water Quality Objectives," which are science-based and equivalent in legal standing to the national "Canadian Water Quality Guidelines," i.e., voluntary guidance benchmarks. Other jurisdictions may identify legally enforceable values as "water quality criteria." International harmonization in terminology would be ideal and practical, but is also likely unobtainable.

Publication

The authoring organization also needs to decide how, when, and how often the water quality benchmarks are published and revised and whether a commitment for upkeep and future development is to be made. Some jurisdictions decided to have a once-only development and publication process, where a group of experts determines a list of contaminants of concern, develops the corresponding water quality benchmarks, and disbands after publication. Other jurisdictions opted for a periodic development and publication process, where a set of benchmarks is developed for the contaminants of concern and published all together every few years (e.g., Australia and New Zealand). As a third option, some jurisdictions have opted for a continuous development and publication process, where new benchmarks are released every few weeks or months (depending on the capacity of the organization, e.g., USA, Canada, etc.). All three options have their advantages and disadvantages, and the authoring organization should decide which is best suited for their purpose.

The "once-only" option has the advantage of being a clearly defined and delineated task, with a predictable time

line and funding requirement. It can respond well to a temporary need for a limited set of specific water quality benchmarks; however, it is not well suited for creating larger quantities of defensible benchmarks. It does not easily allow for future corrections of published benchmarks or for updating the values as more toxicological information on the substances becomes available. Also, a considerable amount of effort will be spent to assemble the necessary expertise to derive scientifically defensible water quality benchmarks, making the creation of only a small set of benchmarks rather uneconomical. Therefore, the "once-only" process can be a good option to obtain a limited set of benchmarks for a particular situation (e.g., to create site-specific benchmarks or to respond to an emergency situation), especially when nestled into an already existing, larger, and ongoing environmental quality benchmark development process.

Both the periodic and the continuous development and publication process require a dedicated team of experts to be at hand to derive new benchmarks, which necessitates ongoing funding and indicates a commitment for the future. While requiring a larger investment of money, manpower, and time than the "once-only" option, the "economy of scale" makes these ongoing processes more economical and will often result in better water quality benchmarks due to the continual growth in experience. If done properly, both have the advantage of consistent quality of work, maintenance, and passing on of acquired expertise in benchmark development (transfer of "corporate knowledge" over the years and consistency in the quality of the derived benchmarks), flexibility to respond to new priorities and emerging contaminants of concern in a timely fashion, as well as to undertake corrections of published benchmarks and the ability to adjust to new information and derivation methods. The "continuous process" will have these advantages probably more so than the "periodic process" as in the continuous development process all stages of benchmark development take place staggered at the same time and "improvements" and experiences gained from one benchmark can continuously be incorporated into ongoing and soon upcoming work of other benchmarks. The continuous process can, therefore, be more flexible and adaptable and can respond quickly to priority changes. However, both are also a "task without end." This open-endedness of the process can be a disadvantage as it brings with it the risk of dwindling or even complete loss of commitment, drive, or funding after the initial euphoria has worn off.

One interesting and potentially time- and resource-saving approach is currently being evaluated in the Netherlands (RIVM 2011; Van Herwijnen et al. 2012). A tiered approach of first deriving an "indicative environmental risk limit" (i.e., ad hoc value) and comparing it with environmental monitoring data before deriving a full-scale water quality benchmark may be employed to keep the workload manageable by allowing to focus on substances with the highest aquatic risk potential.

Implementation

Also, the authoring organization should address the intended interpretation of its benchmarks and provide implementation guidance. This can be in the form of statements like "Our water quality benchmark shall be a threshold level below which adverse effects to aquatic life are not expected. If this benchmark is exceeded, there is an increased probability for an adverse effect to occur." Due to the limitations inherent to the available methods on how water quality benchmarks can be derived (e.g., still limited knowledge in aquatic toxicology and ecology, generally only a small number of species being tested, only some of the possible toxic endpoints and effects being reported, extrapolation of laboratory test results to ecosystem level, species adaptation, ecosystem redundancy and resilience, and contaminant/stressor interactions in the field; see below) and the dosage-based response dependency of toxic effects, their numerical value should not be considered to be an absolute, unalterable limit. Even in the best of cases, a science-based water quality benchmark is a guidance value, indicating the environmental concentration of a contaminant where, based on the best available toxicological information at the time of derivation, toxic effects on aquatic organisms can start to occur, but are not guaranteed to occur yet [only when, of course, the water quality benchmark is designed to be protective and not intended to be an impact indicator as certain special water quality benchmarks (see, e.g., the Canadian short-term exposure guidelines which are designed to estimate severe effects thresholds; CCME 2007a)]. Therefore, rather than being a single value, a water quality benchmark should provide a range, or even several ranges, with associated percentages or levels of effects occurring. This, however, has not yet been done widely [exceptions are Australia and New Zealand (ANZECC and ARMCANZ 2000) and Canada (EC 2013)] and may also only be possible for a few select, well-studied substances. But it could be an interesting challenge for future water quality benchmark developers.

With respect to the implementation of the water quality benchmarks, it is beneficial if the authoring organization also provides guidance on what it means if a benchmark value is exceeded in the ambient environment and what to do in this situation. For example, a very infrequent and only slight exceedance has a different environmental impact potential than many huge exceedances within a short time frame, and these two situations likely require different remediation actions. The benchmark can, for example, be used as a trigger value for further investigation. Above the benchmark level, a further investigation into the severity of the situation and potential measures for emission reductions has to be carried out. Equally, the benchmark can be treated as a trigger for intervention. Then, at this benchmark level, remediating action has to be taken. Of importance here are considerations like size of exceedance, frequency of exceedance, spatial extent of the

exceedances, and location. The range of potential responses depends on the societal values, the jurisdictional framework, the socioeconomic circumstances, and the technological options and is beyond the scope of this paper. Furthermore, it has to be decided what to do if a newly derived benchmark value is below the currently achievable limit of detection for environmental samples. This situation occurs frequently for pesticides and newly identified substances of concern. A jurisdiction may take this limit of detection as the water quality benchmark; however, this is not the best solution. In order for a water quality benchmark to be defensibly protective, it has to be based on toxicological data, and consequently, the benchmark value gives an indication when toxic effects might occur in the environment. Not being able to reliably quantitate a substance in ambient waters is a reflection of the current capability of analytical chemistry for this substance and not a statement on its toxicity in the environment. Therefore, a water quality benchmark below the limit of detection is a call to improve the analytical methods for the substance in question and not a reason to raise the water quality benchmark value. As an interim solution, a temporary water quality benchmark could be set at the limit of detection; however, such a benchmark should be time-limited and accompanied with measures to improve the analytical capabilities.

Site-specific water quality benchmarks

Equally, it has to be recognized that the concentration of a substance in the ambient environment is the result of natural factors, human actions, or a combination of both and that these concentrations change over time and space. Both the natural and anthropogenically caused variations in concentrations over time can be quick (i.e., over hours or days) or slow (i.e., seasonal, decades, centuries), and spatial differences can occur abruptly over very short distances (intercept of two different surface geologies, upstream versus downstream of a significant point source at a river, etc.) or gradual over large areas (along a river with diffuse sources, estuaries, near shore versus open ocean). With respect to naturally occurring substances, it is important to distinguish between the portion of the concentration that is due only to natural causes (i.e., the natural background concentration) and the portion of the concentration that is due, at least in part, to anthropogenic causes. However, quantifying these two portions reliably is often challenging. A water quality benchmark designed to apply over a large geographic area (e.g., a national water quality benchmark) is derived considering all acceptable and applicable toxicological data from a variety of toxicological studies (i.e., including organisms from different aquatic ecosystems and regions and experimental exposure conditions resembling different geological backgrounds). As the natural background concentration of naturally occurring substances is a very site-specific matter, it often cannot be adequately addressed by such a (national) benchmark. It regularly happens that the recommended national benchmark value falls below the natural background concentration (or outside of natural conditions) of a particular site of interest, for example, with many benchmarks for metals applied to mineral-rich areas (as in the vicinity of mining sites). This fact does not invalidate the national benchmark or its derivation process, but it shows the need to understand this derivation process and to know how to properly apply benchmark values. It generally leads to the derivation of site-relevant values (i.e., site-specific water quality benchmarks) to better reflect the adapted local ecosystem.

It is the task of the authoring organization to provide guidance in such a case (i.e., when a recommended benchmark falls below the natural background level). One option can be to recommend that, where the site-specific natural background concentration of a substance exceeds the national benchmark value [derived primarily from generic (non-site-specific) laboratory toxicity data], the natural background concentrations should be taken as the site-specific benchmark value unless or until another appropriate site-specific value is derived according to recommended methods (see, e.g., CCME 2007a). This approach is based on the assumption that the biological community present at a site has adapted to the local conditions, including a naturally elevated level of the substance of concern. It does, however, not imply that the adapted community may be able to adjust to an additional, anthropogenically created exposure to this substance without showing negative effects. This can only be determined with appropriately designed site-specific toxicity studies and can generally not be deduced from generic, non-site-specific studies.

Role of water quality benchmarks

The authoring organization should also address the role of these water quality benchmarks in the jurisdiction in question. A water quality benchmark can fulfill several roles. For example, it can be a tool to evaluate and interpret environmental monitoring data. In this, it becomes an assessment tool to determine the specific or overall ecosystem health and can be an integral part of any reporting on the state of the environment pertinent to the jurisdiction in question. An example for this use of water quality benchmarks is the Canadian Water Quality Index (CCME 2013b), a communication and education tool that summarizes a number of water quality variables into a single measure (i.e., score) of overall water quality. A water quality benchmark can also be a legal tool and serve as the basis for environmental protection and prosecution. Furthermore, they can be starting points to create industrial and municipal release and effluent limits. But under no circumstances should water quality benchmarks be considered as "pollute-up-to permits." The authoring organization, therefore, has the duty to also implement an accompanying "Anti-degradation Policy," i.e., a declaration to maintain (if

pristine) and improve (if already degraded) the existing quality of the water bodies. As this has to be balanced with desired resource development, policy decisions to allow and to define "acceptable levels of impact" are required. Examples of antidegradation policies can be found on many jurisdictional web sites, water strategy plans, and water quality benchmark guidance documents (e.g., US EPA 2013; CCME 2003, 1999).

Various forms of water quality benchmarks

As described above, a water quality benchmark is generally a threshold level(s) and guidance value(s) which aims to approximate the level where there are no observable effects (or accepted effects) to aquatic life. Such a benchmark can either be quite simple (e.g., a single value), or be more complex (e.g., a range or values), or quite extensive (e.g., equations, tables, or matrices). It can be a numeric value(s), a narrative statement, or a combination of both. The former option is used mostly for chemical substances, while the latter is used often for environmental parameters (such as pH, temperature, turbidity, water hardness, etc.).

Prioritization of substances of concern

Another necessary aspect to be considered by the authoring organization is a Prioritization Scheme to assist in the identification, validation, and ranking of the substances (contaminants of concern) that are of importance for water quality benchmark development in the jurisdiction in question. The scheme should be an integral part of the periodic overall priority setting and work planning of the authoring organization. Many jurisdictions have such a prioritization scheme, either formal or informal, customized for their needs (e.g., Article 16 of the Water Framework Directive; European Communities 2000; Environment Agency 2007; CCME 1999, 2007b).

When creating a ranked list of substances for which there is a current priority need for benchmark development, the scheme should take into account the science, policy, and regulatory drivers, internal and external consultation, as well as the underlying scientific feasibility (i.e., toxicological data availability) of benchmark development. A priority list should be updated periodically in order to accommodate shifting priorities and emerging issues. For this to work, the required fact finding (i.e., scoping), consultation, and priority setting processes must not be too elaborate and time-consuming. Short turnaround times are essential, and incomplete information must be accepted. Consequently, the resulting list of substances must be considered as a "draft list" (or "evergreen list") to be periodically revised as required and consulted for guidance purposes only. As it is impossible to periodically evaluate a large number of substances in a short time period for priority setting, the scientists and regulators involved should exercise their professional expertise, experience, and judgment and utilize outside resources, where available (e.g., other jurisdictions, environmental non-governing organizations, industrial stakeholders, and others) in developing a short list of approximately 20–30 substances to be evaluated under the prioritization scheme. The scheme will then validate (i.e., support or reject) and rank the substances on this short list. Such a process is used, for example, in the prioritization of the substances intended for Canadian Water Quality Guideline development (CCME 2007b).

The prioritization scheme can use a set of questions (addressing science- as well as regulatory- and policy-related aspects) to provide "points" for the different substances. The individual points for a substance are tallied up, and if a substance scores above a minimum number of points, it is considered a priority for benchmark development. While these points give an indication of the comparative relevance, professional judgment should also be used in the final decision in developing a benchmark for a substance identified as a priority (CCME 2007b).

Targeted, directed consultation within internal and external stakeholders is an important part in determining the priority status of a substance. It should be a planned, result-driven communication with identified stakeholders that have direct and vested interests in the growth of water quality benchmark development. The prioritization will also benefit from their expertise and experience. The purpose of the consultation is to determine and understand the needs of the stakeholders and to ensure that the benchmark development program can respond to these needs. As such, this consultation should iteratively proceed from the general to the specific, i.e., from an overall scoping and search for candidate substances to the verification of the priority of the selected substances.

Summary

The authoring organization has the important task of setting up and providing guidance throughout the water quality benchmark development process in accordance with the particular jurisdictional requirements. In this task, many science-based as well as policy-based decisions must be taken. A thorough understanding of the issues involved, good planning, and prudent guidance will result in a successful water quality benchmark development program.

The derivation protocol

A derivation protocol is a guidebook on how to derive the water quality benchmarks for a particular jurisdiction (see, e.g., CCME 2007a; Lepper 2005; ANZECC and ARMCANZ 2000; Stephan et al. 1985). It outlines and incorporates many of the same aspects stated above by the authoring organization, such as the purpose, protection level, application area, etc., of the water quality

benchmarks. However, it also goes beyond and into much more detail, especially on the technical side, such as guidance on acceptable and unacceptable toxicity information and the actual methodology of deriving the benchmark. The protocol describes the separate development steps necessary to create a benchmark value from first identification of substance to last approval and publication. By giving detailed guidance, it provides not only consistency in the derivation of benchmarks for all substances of concern but also scientific credibility, transparency, and defensibility to the process and to the individual benchmark values. The derivation protocol is a valuable tool in communicating the purpose and use of the water quality benchmarks.

It is important to realize that the derivation methods of long-standing jurisdictions [e.g., USA (Stephan et al. 1985) and Canada (CCME 1991), among others] were generally developed decades ago and were based on the available science of this time. Some or many aspects of these old derivation methods may be outdated as our understanding of aquatic toxicology and methodology has advanced. They were designed to work within the jurisdictional framework of their country at the time and were based on the aquatic ecology of their geographic area. More recently implemented derivation methods, for example many of the new Asian methods (Wu et al. 2010; An et al. 2011; WEPA 2012; Feng et al. 2012), are often based on these established methods, sometimes even without recognition or consideration of their limitations or newer scientific understanding (such as shortcomings of NOEC/LOEC endpoints, issues with safety factors, bioavailability of metals, secondary poisoning, endocrine disrupting ability, etc.). Therefore, when developing a new jurisdictional program, an existing method should not be adopted or transferred without a thorough analysis and understanding of all relevant aspects, and updating where necessary. Equally, aquatic toxicology is an ever-changing and evolving field; therefore, jurisdictional approaches to manage environmental issues must be flexible and adaptive, and consequently, derivation methods for water quality benchmarks should be periodically reviewed to incorporate new tools.

The Canadian case study

The actual development of the derivation protocol can sometimes take very little time and money. For example, the first Canadian protocol (CCME 1991) was created in less than 1 year, but this was only possible because the protocol was written after the fact; that is, in 1991, Canada already had a well-established national water quality guideline development program in place since 1984 and comparable federal and provincial programs even before this date. At the time of drafting the protocol, several hundred water quality guidelines had already been published, and the task became merely a retrospective analysis of what had been done for years. Still, under regular circumstances, the development of a new protocol will take a lot more time and effort, as was shown by the second Canadian protocol (CCME 2007a), which took about 7 years to complete. In 2000, the authoring organization, i.e., the "Water Quality Guidelines Task Group," decided that it was time to update the existing derivation protocol and incorporate new scientific methods. Under the jurisdictional framework of Canada (Canada is a federation of independent provinces and territories), a multi-jurisdictional protocol development group was formed with the task of creating a new guidance document. This group was asked to explore new derivation methods while maintaining certain well-working aspects of the former protocol and addressing differing jurisdictional needs and requirements. The new Canadian protocol development process became a lengthy and tedious journey. This came about not only because of the goal of going beyond existing methods (which necessitated time-consuming education, analysis, testing, and verification of newly emerging methods) but also because of a historical oversight of several of the duties of an authoring organization outlined above (which required lengthy multi-jurisdictional consultation to rectify). But it ultimately turned into a rewarding and insightful endeavor.

Some of the key experiences gained in this process are presented in this paper in the hope that other jurisdictions will benefit.

Insights gained

One of the most interesting observations was the desire to validate the new derivation methods by comparing the new benchmarks to the results from the traditional derivation methods. While there was unanimous agreement that new derivation methods should be utilized [as the limitations of the existing derivation method (lowest acceptable endpoint multiplied by a safety factor) were understood], there was also great hesitation to trust the new benchmarks as being "protective enough" without comparing their values to the results from the traditional method. Even though it was realized that the traditional method sometimes yielded questionable guideline values, and at times even unacceptable and scientifically indefensible values, some players were reluctant to accept a new method without a detailed analysis and comparison of the results to the traditional method for several sample substances. The desire to use "new science" to derive water quality benchmarks was often tempered with the inertia to break with the traditional and to venture outside the established "comfort zone." This cautionary approach was also partly due to a gradient in the knowledge and understanding of the technical details of benchmark development by the different players involved. Bringing everyone up to the same level of knowledge necessitated not only knowledge transfer within the inner protocol development group and with the larger authoring organizational group but also extensive consultation with

outside stakeholders and outside experts. While this resulted in an increase in time and cost, it also resulted in a much deeper understanding of the advantages, disadvantages, problems, and potential errors associated with the various new methods that were explored. In the end, it likely yielded a better product.

Another delaying aspect was a change in scope midway into the project. The project started as an addendum to the existing derivation protocol to "simply" correct existing inadequacies with respect to a group of contaminants of concern (i.e., metals). But as the realization into the depth of these existing inadequacies and the need to update the whole protocol grew closer to the completion of this narrower task, the scope was expanded to revise the whole existing protocol and apply it to all substances.

Other challenges were the differences in the Federal and Provincial jurisdictional needs and differences in the current understanding and interpretation of the traditional protection goal and the associated level of protection of the Canadian Water Quality Guidelines. While this goal and the level of protection had originally been stated in brief in 1987 (CCREM 1987), over time, jurisdictions had developed different understandings and interpretations on how to achieve this goal (e.g., protect ecosystem function, protect individual species, or even protect individual members of a species, i.e., allow limited impact to species or individuals, but maintain ecosystem function versus disallowing any impact to occur on species or individuum level). In the absence of detailed historical records to refer back to, a new consensus had to be worked out.

A major challenge was the logistical aspect associated with the project. Members of the protocol development group came from various Canadian Provincial and the Federal Ministries of the Environment and lived and worked thousands of kilometers apart. Competing commitments and differing levels of support by the respective jurisdictions often made active participation in pending issues and attending meetings challenging for many participants. While a lot was accomplished by e-mail, individual or bilateral preparatory work, and via teleconferences, actual face-to-face multiday workshops and meetings proved to be invaluable and indispensable, albeit still difficult to organize.

In summary, the development of a new derivation protocol for the Canadian Water Quality Guidelines became an interesting but worthy journey with many unexpected twists and turns, surprises, and rewards. Among other valuable results, it yielded a deep understanding of the various derivation methods, including the species sensitivity distribution approach (see below).

Benefits of a team approach

As challenging as it was in developing a water quality benchmark derivation protocol, it became quite clear that such a task would require a team approach. There are too many aspects to consider and which require specific expertise in different subject areas for one person alone (ecology, toxicology, chemistry, statistics, etc.). Furthermore, new ideas are created and are improved through discussion and challenge within an expert group. Still, it requires that the team members be open-minded, dedicated, and committed to the process, but also that they are already experts with hands-on experience in benchmark development in order to draw from their own experience when evaluating new approaches. Team members also need to be flexible and must understand consensus, especially when working within a multi-jurisdictional program. As there is a requirement of different areas of expertise, team members must be able to lead parts of the project pertaining to their expertise. Therefore, the team should consist of a leader who can follow and followers who can lead. Everyone involved must be a team player. The members must be available (i.e., must be able to dedicate time and be able to travel). This requires that the authoring organization secures sufficient funds and time to complete the project successfully. It must be realized that it will take time, money, commitment, and fortitude.

Issues to address in a derivation protocol

As described earlier, the derivation protocol is the guidebook for the water quality benchmark development process. It clarifies many technical and scientific details that have to be considered in order to obtain a scientifically defensible and environmentally protective benchmark that responds and fulfils the particular needs and requirements of the authoring organization. Therefore, a derivation protocol has to give guidance on the following issues.

Determination of acceptable studies

Aquatic toxicity studies are performed for many reasons and different purposes, with different testing methods, studying various endpoints and impacts, and yielding a range of results. Some are performed as part of exploratory scoping, some are set up to test a battery of substances or organisms, and some are done to test a single variable (e.g., substance, organism, or endpoint) in detail. Some follow established testing protocols and others use new experimental methods. Some were done with the greatest care and yield reliable results, while in others the experimenters had made mistakes that put the results in question. Therefore, a great deal of variability exists in the quality of published toxicity data. Not all published and unpublished toxicity studies are suitable for use in benchmark derivation. Consequently, every available study should be scrutinized in detail and evaluated for its suitability. Guidance on proper evaluation of toxicological studies is abundant and can be found in already published water quality

benchmark derivation protocols and numerous publications. Examples are Ågerstrand et al. (2011), ANZECC and ARMCANZ 2000; CCME 2007a; ECHA (2012), European Communities (2011), Klimisch et al. (1997), Lepper (2005), Mensink et al. (2008), and OECD (2005).

As there is a great variation in potentially useful toxicity studies, their evaluation should not follow a rigidly fixed format but rather should allow for special consideration and incorporate scientific judgment on a case-by-case analysis. It is not necessary that a study follows a standard test protocol; non-standard testing procedures should be evaluated on their own merit as they may yield results usable for guideline development. But the study design and execution must be appropriate with respect to the test substance and organism. For example, while flow-through tests are generally preferred over static renewal tests, in both, the volatility of a substance must be appropriately addressed. Some substances adsorb to the walls of the test container and tubing or degrade quickly, thereby reducing the actual exposure concentration. These effects are some of the reasons why tests with nominal concentrations are usually of questionable quality. Others are potential mistakes in the preparation of the stock solution or the serial dilutions. Equally, the solubility limit of the substance in relation to the tested concentrations is important to consider, as well as the chemical behavior and potential interactions of the test substance with other chemicals present [pH, dissolved oxygen, salinity, organic matter, adjuvants (chelators), carrier solvents, hardness (Ca^{2+} and Mg^{2+}), and alkalinity]. The actual tested concentrations should be rather closely spaced and evenly distributed around the effect concentration; therefore, scoping or range-finding tests may be deemed not suitable or acceptable for inclusion in datasets for benchmark derivation.

Control studies must be done and the number of test organisms and of replicates must be appropriate to allow for adequate statistical analysis of the results. The statistics used in evaluating the data must be described and be suitable to the data. The analytical methodologies to monitor the test conditions and measure the actual test concentrations must be appropriate.

The experimental test conditions must be appropriate for the tested organism (e.g., with respect to temperature, pH, hardness, salinity, etc.) and potential toxic interactions with other substances [especially adjuvants (chelators), and carrier solvents] must be considered. The experimental design must fit the organism's requirements. The test organisms must have been adequately acclimatized to the laboratory conditions.

Impact of cryptic and adapted species

Related to this is the growing understanding of the existence of cryptic species; that is, a seemingly uniform population has split into two or more species, sometimes even occupying (or partially occupying) the same habitat (Pfenninger and Schwenk 2007; Gabaldón et al. 2013; Hogg et al. 1998). These two species seem identical, but are already genetically distinct and differ in their sensitivity to toxic effects of substances (Feckler et al. 2012; Rocha-Olivares et al. 2004; Duan et al. 1997). Not recognizing them as separate species can have implications on the toxicity evaluation and subsequent benchmark derivation, depending on the particular requirements of the derivation process (e.g., with respect to fulfilling a minimum data requirement, inappropriate averaging of multiple results for a species, representation in a species sensitivity distribution curve, etc.).

Equally, consideration must be given on how to treat the toxicity test results from differently adapted populations of the same species. Especially when studying naturally occurring substances (e.g., metals) and using field-obtained organisms, special attention must be given to potential prior exposure and adaptation of the test organisms to the test substance. Organisms obtained from areas with naturally or anthropogenically elevated levels of the test substance have likely adapted to these higher levels and are generally less sensitive (more tolerant) to its toxic effects. However, they may also be already more stressed than organisms from other areas and, as a result, less tolerant (i.e., more sensitive) to this or other substances (e.g., a Daphnid species strain originating from naturally metal-enriched areas versus a Daphnid species strain from naturally metal-deficient areas). They will respond differently in their sensitivity to certain substances (Barata et al. 2012; Nys et al. 2012).

Neither adapted nor cryptic species have been addressed in any of the published protocols for water quality benchmark derivation or environmental risk assessments, nor are there any actual water quality benchmarks published yet where these issues have been considered. As genetic fingerprinting is becoming cheaper, easier, and more common, it is now prudent to include the genetic fingerprint and appropriate details on the source and origin of the test organism in the reportable information of a toxicity test. Equally, consideration of the potential occurrence of cryptic and adapted species should be given in future derivations of water quality benchmarks.

Determination of the acceptable toxicity estimator

In aquatic toxicology, there are basically two different types of toxicity estimators or toxicity measures obtained through the statistical analysis of the toxicity test results, i.e., hypothesis-based and regression-based estimators. The traditional estimators were obtained through hypothesis-based statistical data evaluation [i.e., no-observed-effect-concentration (NOEC) and lowest-observed-effect-concentration (LOEC) values]. Even though these two estimators have been used extensively in water quality benchmark derivation, they should no longer

be used here as they are generally unsuitable for this task, unless the underlying experiment was executed exceptionally well (e.g., Hoekstra and Van Ewijk 1993; Chapman et al. 1995, 1996; Warne et al. 2008). Their names are misleading and false. Their names imply that they are the concentration for a toxic substance that elicits no observable toxic effect and the lowest concentration that elicits an observed effect, respectively. Neither is true. The NOEC is the highest test concentration where the observed effect is not statistically different from the effect observed at the control concentration; the LOEC is the lowest test concentration where the observed harmful effect is statistically different from the effect observed at the control concentration (Rand 1995). By definition, the LOEC is the next higher tested concentration of the NOEC. The important points to realize here are that (a) the concentrations are test concentrations, i.e., they are dependent on how well or poorly the test is designed; (b) the impact and difference assessed is statistical and not toxicological; and (c) the magnitude (i.e., toxic impact) of the observed effect is irrelevant in the determination of the NOEC and LOEC. In well-designed and well-conducted toxicity tests, this may not be a problem, and they may turn out to be close to or even at the intended concentrations (that is, the interpolated concentrations that would still not elicit a harmful effect and would start to elicit a small harmful effect, respectively). However, in reality, for most tests, this is not the case (Moore and Caux 1997). Often, the NOEC already causes noticeable and measurable harmful effects to the test organisms and the LOEC causes considerable damage, sometimes even approaching EC_{50} or LC_{50} impact levels; that is, they are not what is claimed they are. For these reasons, the use of NOEC and LOEC (and similarly derived) estimators is not acceptable in the derivation of water quality benchmarks that are designed to be protective and scientifically defensible.

Unfortunately, NOECs and LOECs are still widely published and often make up the bulk of the available toxicity information for a substance. Not using them will prevent the derivation of a water quality benchmark for many substances. Therefore, the authoring organization and the benchmark-deriving experts must have a clear understanding of the risks and impacts associated with the use of these estimators, and the derivation protocol must provide clear instructions whether and how to use them. Their use should be diminished and, over time, eliminated completely.

More suitable for benchmark derivation are the toxicity estimators that are obtained through regression-based statistical analysis of the toxicity test results [e.g., no-effect-concentration and effect-concentration (EC_x, at x% level), etc.] as this analysis allows identifying values that represent no- or low-effects thresholds. Strictly speaking, only the EC_0 is a true no-effect level; anything higher (i.e., EC_{01}–EC_{10}) is a low-effect level. But in order to accommodate for experimental variability, the threshold level for no negative effect is sometimes defined as an effect level on 10 % or less of the exposed individuals of a species (i.e., EC_0–EC_{10}) and, similarly, a threshold level for low effects as an effect level on 15–20 % of the exposed individuals of a species (i.e., EC_{15}–EC_{20}; as, for example, in CCME 2007a). Each authoring organization must clarify its definition. Sometimes, a more appropriate no-effect or low-effect threshold is identified for the test species in a generally accepted standardized test protocol (i.e., the most appropriate EC_x representing a no-effect threshold for the species).

It happens sometimes in toxicity tests that an insufficient concentration range on the higher end has been tested and the results have to be expressed as "toxic concentration is greater than x." While it is not desirable, the use of this data is generally acceptable as it will not result in an underprotective benchmark. Such data are best used as supporting evidence for other studies and to help fill minimum data requirements. But such a study must be evaluated thoroughly as there is a good chance that the study may not be acceptable for other reasons. Consideration must also be given on how many such data points should be included in the benchmark derivation, i.e., the percentage of "greater than" data points compared to the whole dataset and how these values compare to the rest of the data. One aspect to evaluate is the assumption that the tested organism is truly tolerant toward the tested substance. The toxic threshold for this organism must be clearly above the identified thresholds for other more sensitive organisms. Substances and situations where this may occur are toxicity studies for pesticides on non-target organisms (i.e., plant toxicity studies with insecticides or pesticides or vertebrate/invertebrate toxicity studies with herbicides). Guidance on the use of such studies should be included in the derivation protocol.

The opposite situation applies to toxicity tests where an insufficient concentration range on the lower end has been tested (i.e., where the results are expressed as "toxic concentration is less than x"). As the true sensitivity (toxicity threshold) of the species toward the tested substance is not known, but is below the lowest tested concentration, these results are definitely not acceptable for benchmark derivation as they will result in an under-protective benchmark.

The authoring organization and/or the benchmark-deriving experts must define in their derivation protocol which of these toxicity estimators (hypothesis-based/regression-based, >/<) should be used in their jurisdiction.

Selection of effect endpoints and extent of extrapolation to benchmark

Extrapolation from the known (i.e., the concentration in a toxicity test that causes a negative effect) to the unknown (i.e., the concentration deemed not to cause a negative effect) is unavoidable in water quality benchmark derivation.

However, the extent of this extrapolation influences the confidence on the protectiveness of the benchmark. Extrapolation from severe effects endpoints (e.g., lethality to a considerable percentage of a population, i.e., LC_{50} values) toward a safe threshold (i.e., concentration not causing any negative impact) is a great leap of faith. Large uncertainties are generally associated with this extrapolation, and consequently, a benchmark derived in this manner is open to criticism and challenges. First, there is a low confidence in the assumption that the benchmark is protective; second, critics often claim that the benchmark is either over- or under-protective, depending on what suits them better. In order to increase this confidence, it is better to extrapolate a protection benchmark from less severe and non-lethal and, therefore, more sensitive endpoints. Endpoints of choice are growth, reproduction, hatching, or germination success, effects on embryonic development, survival of juvenile stages, etc., but can include behavioral (e.g., avoidance behavior, mating rituals, nest site selection, migration, etc.), fitness-related, hormonal (e.g., endocrine-disrupting), pathological, or physiological effects, especially if it can be shown that these effects are a result of exposure to the substance in question, lead to an ecologically relevant negative impact, and the tests are scientifically defensible.

Similarly, the confidence of having derived a benchmark that is protective is increased if the extrapolation is small and done from lower effect levels [i.e., no effect or low effect (e.g., EC_{10} or EC_{20})] rather than severe levels (i.e., EC_{50} or LC_{50}) and where the test organisms were exposed to the substance over longer periods of time. This applies independent of the benchmark derivation method (e.g., most sensitive study multiplied by a safety factor, species sensitivity distribution, or other approach). A protection benchmark derived from 96-h LC_{50} values requires larger extrapolations and has less confidence in being "on the mark" than a benchmark derived from 21-day EC_{10} values for growth or reproduction impairment.

Sometimes, jurisdictions derive impact-indicating benchmarks, i.e., levels where it is fairly certain that environmental impacts will occur. Examples are the Canadian short-term exposure guidelines (CCME 2007a) or the USEPA acute criteria (Stephan et al. 1985). These are not protection benchmarks, but rather triggers for immediate action (e.g., spill cleanup, prosecution, etc.). Here, confidence that effects will occur at the benchmark level is desired, and consequently, derivation is based on severe effect levels (e.g., 96-h LC_{50} values) with no or only minor extrapolation.

Safety factors

Safety factors are often used in water quality benchmark derivation in the extrapolation from the known to the unknown (see above). Their use, magnitude, and associated issues, etc., have been extensively discussed (e.g., Pohl et al. 2010; Malkiewicz et al. 2009; Dourson 2005; Elmegaard and Jagers op Akkerhuis 2000; Chapman et al. 1998; Pieters et al. 1998; Renwick 1995) and are beyond the scope of this paper. But as their use in benchmark derivation seems unavoidable, especially for substances with little toxicological data, an authoring organization and the protocol derivation experts are advised to familiarize themselves with these issues in order to make appropriate science- and policy-related decisions.

Use of corroborating data

The authoring organization must also decide whether microcosm, mesocosm, and field studies are acceptable in their derivation process and, if so, under what conditions. One requirement for acceptance should be the existence of a dose–response relationship in the results and a reasonably defensible apportionment of the effects to the substance. While this may be possible for micro- and mesocosm studies, field studies generally have too many uncontrollable and recordable variables, and it is unlikely that they can be used in benchmark derivation. However, while not directly contributing to the actual value derivation, they can play a significant role in evaluating and validating toxicological endpoints obtained in the laboratory and can corroborate a water quality benchmark.

Biological data requirement

Clear guidance must be provided in the protocol on the biological data requirements; that is, what kind and how many studies on specific organisms and under specific conditions (especially which exposure durations) are required to proceed with the derivation of a benchmark. Also, it must be clarified which species can be used, that is, all species (globally) or only native (or endemic) species. The goal of such a minimum data requirement is to ensure that (a) the derivation method is robust, especially if it involves statistical analysis; (b) toxicity information on a reasonable amount of species is available and considered to at least approach some resemblance of ecosystem coverage; and (c) the resulting benchmark is applicable to the intended area of implementation. The details will depend on the derivation process and policy drivers, but the authoring organization and the protocol derivation experts must balance the understandable desire for more data on more species with the unfortunate reality of a general paucity of data for nearly all substances of concern. Absence of information should not unduly prevent the derivation of a needed water quality benchmark. This is why some jurisdictions [e.g., Canada (CCME 1991; CCME 2007a) and Australia/New Zealand (ANZECC and ARMCANZ 2000)] have chosen to derive a tiered set of benchmarks, with different minimum data requirements.

Depending on the complexity of the derivation method(s) selected, it may also be required to give guidance on the classification of the studies and the ranking and preferential use of endpoints.

Final verification

Instructions on the final verification of a benchmark value are also beneficial. This final verification includes a comparison of the derived benchmark value to toxicity information that was not used in its derivation. If the benchmark was derived using LC_{50} (because they are abundant and robust), it has to be compared and validated to other more sensitive endpoints to verify that it is protective. Conversely, if the benchmark was derived using, e.g., EC_{10} values for sensitive effects, it has to be compared to endpoints (including LC_{50} values) for species which were not represented in the EC_{10} dataset. As short-term lethality tests are the most commonly performed toxicity test, it is possible that only a LC_{50} value but no EC_{10} values exist for a sensitive species and that this LC_{50} value is actually close to or even lower than the derived benchmark value. In this case, the proposed benchmark value is likely not protective and an alternative derivation method may be more suitable.

Flexible guidance

It is important to realize that in this process, science can inform and guide, but it can never resolve all decision points. Many decisions related to water quality benchmark derivation are policy-based and not science-based. Also, not all substances will fit into a single derivation scheme, and not all situations and exceptions can be foreseen. The derivation protocol should provide as much guidance as possible without being overly prescriptive, but it also needs to be flexible and allow for exceptions.

Derivation of water quality benchmarks

It goes without saying that a water quality benchmark should be scientifically sound and defensible. And it should be based on relevant aquatic toxicity data. Only in exceptional circumstances, such as absence of toxicity data, should a benchmark be derived using other non-toxicity-based methods.

The composition of aquatic plants and animals and various physiological processes vary naturally with the physical, chemical, geological, and hydrological conditions of the local environment. Water quality benchmarks can, therefore, be designed to be applied to these differing freshwater, estuarine, or marine ecosystems and for arctic, temperate, and tropical conditions. While most national benchmarks are broadly applicable, regional or site-specific benchmarks have a more limited geographical application area. These application areas determine the chemical, toxicological, and ecological data requirements (e.g., aquatic toxicity information on tropical organisms versus arctic organisms, chemical fate in warm waters versus cold waters, etc.).

A water quality benchmark development process has three distinct phases:

- Project planning and initiation (substance selection, funding allocation, work planning, etc.)
- Technical execution (i.e., benchmark derivation)
- External interaction (i.e., review, approval, publication)

And each phase has important steps which need to be followed.

Project planning and initiation

During the project planning and initiation phase, the substance or parameter of concern for which a benchmark is to be developed is identified. This should be done according to the previously mentioned priority setting process. Key associated decisions and tasks are the securing of the necessary funding for the project, work planning and outlining the time lines, identifying the relevant experts and stakeholders (both for drafting and for review), and assembling the technical expert team to draft the benchmark.

Technical execution

The second phase, the actual technical execution of the benchmark derivation, follows the process outlined in the derivation protocol. The first step in this process is the gathering of all relevant and essential information. This may be limited to assembling existing information through literature searches and contacting stakeholders and other jurisdictions for relevant data and documents, but may also include the generation of new data through targeted scientific studies (i.e., toxicity studies, environmental fate and behavior studies, chemical degradation studies, etc.) when required information is not available. The thoroughness of how this information gathering is done will greatly influence the final quality and scientific defensibility of the water quality benchmark. Shortcuts taken in this process may lead to extensive and time-consuming iterations in the review stage. However, the trade-off between undertaking all the necessary studies to populate a fully comprehensive benchmark document and the timeliness of releasing a technically defensible benchmark need to be considered as well.

The relevant and essential information about a substance is not only toxicological data. The physical and chemical behavior (i.e., fate and pathways) of a substance in the aquatic environment of concern must also be understood when deriving a water quality benchmark. It is, however, not necessary to have complete understanding on all aspects; the goal should be to produce a general assessment. Desirable information includes, but is not limited to, the solubility of the substance of concern in the various relevant aquatic systems (e.g., hard/soft and acidic/alkaline freshwater, estuarine, marine, arctic to

tropical conditions); the mobility of the substance in the aquatic environment; migration in and out of sediment and air; potential chemical reactions and the eventual chemical form under various environmental conditions; and the persistence in different media (water, sediment, biota, as well as soil and air, where relevant). Related to the persistence of the substance is information about the chemical breakdown and creation of by-products as they may still be toxic and should, therefore, be incorporated in a benchmark. In addition, data on the ambient environmental concentrations in areas of relevance (and, potentially, areas for comparison) and, where applicable and possible, information on whether elevated levels are due to natural or anthropogenic causes, as well as analytical and toxicological testing methods (including current detection limits), will be useful.

With respect to biological and toxicological information, an understanding of the potential routes of exposure and uptake by aquatic organisms, the mode of toxic action and related toxicokinetics, the bioavailability and the conditions under which the substance is bioavailable, its bioaccumulation potential, metabolic essentiality (if applicable), and toxic interactions with other variables and behavior in mixtures is required. Essential is, of course, all relevant data on the actual toxicity to aquatic biota after long-term exposures (and, where applicable, short-term exposures) with respect to the different effects (e.g., lethality, impacts on growth, reproduction, survival fitness, behavior, etc.).

Of additional benefit for comparative analysis, and as potential source of information, are existing water quality benchmarks from other jurisdictions. As described above, it is important to understand the origin and context of these benchmarks, and any meaningful comparison should be more than just a superficial juxtaposition of the numerical value of the various benchmarks.

The second step is the actual evaluation of the assembled information. During this evaluation stage, the available information is assessed for suitability and acceptability. In order to be able to create a defensible benchmark, the information used in its derivation has to be of acceptable quality. Especially for the toxicological information, this requires that the original source of the information is consulted and thoroughly evaluated, i.e., the actual scientific study rather than a secondary compilation (i.e., database, summary document, textbook). Critically examining the available knowledge on the environmental fate and behavior of the substance is as important as the detailed evaluation of each individual toxicity study. The goal at this stage will be a detailed understanding of the behavior of the substance in the aquatic environment under relevant ambient conditions, its interactions with other substances and factors, and its toxic potential to aquatic organisms. It should come as no surprise that each substance will pose its own challenges: sparsely studied compounds due to the scarcity of information and well-studied substances due to the myriad of conflicting information, and not all of it of acceptable scientific quality. The task of the benchmark developer is to critically determine the scientific reliability of each individual bit of information and piece it together to obtain an overall picture. This requires a careful examination of all relevant studies, with the intention of not only detecting any potential mistakes or shortcomings (i.e., reject a study for use in the benchmark development) but also validating the results and conclusions of a study (i.e., prove that the study is acceptable for use in benchmark development). Once all the relevant studies pertaining to a specific issue have been examined in detail, it is often possible to resolve conflicting information and contradictory results. Many of the published derivation protocols give good guidance on how to perform such an analysis.

After all available information has been assembled and assessed, and unacceptable study results have been excluded, it is now possible to proceed to the third step, the derivation of the actual water quality benchmark value. This process will follow the method(s) outlined in the derivation protocol [e.g., lowest acceptable endpoint multiplied with safety factor, statistical extrapolation (SSD or similar), or other potential methods]. The method selected will depend upon the fulfillment of the requirements (especially the minimum toxicological data requirements) stipulated in the protocol.

The fourth step will be the drafting of the body of the text of the benchmark. The length and actual content will depend on the available information on a substance and the outline given by the derivation protocol, but at the very least, the document should present the relevant information (supporting data and toxicological data) required to understand and properly apply the water quality benchmark. This includes pertinent environmental fate and behavior as well as aquatic toxicity results. It is important to not only present the data and results that are used, i.e., deemed relevant and acceptable, but also to discuss and explain unaccepted and rejected results. Presenting this information will serve two purposes. Firstly, if rejected results are not mentioned, it is likely that during the peer review stage of the document they will be flagged as missing information, which creates doubt on how thorough steps 1 and 2 have been executed and questions the defensibility of the benchmark value. Presenting this information aids and simplifies the review process and preempts unnecessary iterations of rewriting and reviewing the document. The second purpose of presenting unaccepted and rejected studies is to inform the greater benchmark developing community, especially colleagues in other jurisdictions, to avoid perpetuation of mistakes.

The water quality benchmark document is a record on the process and reasoning followed in the derivation of the actual benchmark. While the derivation usually follows the process outlined in the protocol, exceptions can occur, but should be explained. Transparency in every step is crucial.

As a thorough and well-written water quality benchmark document can become quite lengthy, it may be advantageous to create an additional, shortened companion document (i.e., a fact sheet) which contains a summary of the essential points of the longer document. This approach has been taken, for example, in the publication of the Canadian Water Quality Guidelines, where generally both a longer and detailed guideline document and a short two- to five-page-long fact sheet is published (CCME 1999).

External interaction

The third phase of water quality benchmark development involves the external interactions, i.e., the review, approvals, and publication steps. The review process includes the examination of the benchmark document by external experts (including benchmark developers from other jurisdictions), relevant stakeholders, and the public, as appropriate and outlined by the authoring organization in the overall process, and is generally an iterative procedure involving often several cycles of rewriting and reviewing. After successful review, the water quality benchmark is being approved by the appropriate authority and finally published. Publication is generally the duty of the authoring organization and may be in the form of a paper in an appropriate scientific journal, but more likely it will be as a document issued by the authoring or higher-level organization (e.g., jurisdictional government, international organization, etc.). In this case, it can be published in electronic and/or printed form, either as a stand-alone document or as part of a broader benchmark compendium.

As a water quality benchmark is generally developed by an authoring organization for a specific purpose and customized to local environmental conditions, transfer and adoption for use in different environments or in other jurisdictions is usually not recommended. But it can, in some instances, be a viable option, especially as an interim solution, when the origin and background of the benchmark has been examined and deemed acceptable.

Species sensitivity distribution as a benchmark derivation method

Over the past few years, the use of species sensitivity distributions (SSDs) in water quality benchmark derivation has increased, influenced especially by the publication of the Australian and New Zealand protocol (ANZECC and ARMCANZ 2000; Posthuma et al. 2002). While the method is simple in concept, it is challenging in detail, especially when being used in benchmark derivation. An authoring organization planning to set up a water quality benchmark program and develop its derivation protocol is therefore well advised to encourage its water quality benchmark-developing experts to examine, understand, and decide upfront crucial points (explained below) related to the SSD approach. Oversight or omission of these points can result in inadequate or erroneous application of the SSD approach.

In brief, the SSD approach entails plotting the available toxicity data for different species as a cumulative frequency plot against concentration, fitting a statistical distribution to it (i.e., the curve fitting), and then calculating the concentration that should theoretically protect any chosen percentage of species [usually the 5th percentile intercept with the *Y*-axis, the HC_5 (hazard concentration)]. However, for use in water quality benchmark derivation, several points have to be considered in order to obtain the appropriate and desired result.

Data quantity

The quantity of the plotted data points determines the robustness (i.e., the stability) of the curve and the degree to which the selected *Y*-intercept is influenced with the addition or deletion of data points. Simplistically said, the more data, the better the fit of the curve. Therefore, the goal is to have as many data points as possible. Depending on the uniformity of the data points, a curve generally becomes fairly stable with 20 or more points. However, even in fairly large datasets, the HC_5 value can change dramatically with the addition of extremely low data. Conversely, even in relatively small datasets, the inclusion of additional data points, especially in the mid-range, may not change the curve and the HC_5 value much. Absence of data on the higher end (i.e., toxicity data for tolerant species) tends to raise the HC_5 value, while the addition of more data on this end tends to lower the HC_5 value. While this influence seems counterintuitive, it is important to keep in mind when creating the dataset for plotting.

While it is possible to reasonably fit a statistical distribution to as few as five or six data points, it is not recommended to set a fixed lower limit based on statistical requirements. It is better to use a well-designed toxicological minimum data requirement [stipulating data for at least six to eight different species (keeping in mind the caveat explained earlier under "Biological data requirement," i.e., general paucity of suitable toxicity information for most substances)] as this will generally also result in datasets that will allow the generation of fairly robust SSD curves. However, in addition, it is advisable to implement statistical "goodness-of-fit" assessment parameters as acceptance limits. But it must be kept in mind that almost all models fitted to small datasets will pass goodness-of-fit tests and that these tests gain strength with more data points. Therefore, a simple pass/fail analysis of a statistical test may not be sufficient, and a more detailed examination of the test result(s) is advised. And it must be pointed out that fairly reliable water quality benchmarks can be obtained with approximately 12–15 data points.

Dataset selection and curve interpretation

It is also very important to have a clear understanding on which endpoints and effect levels to plot as this will determine the interpretation of the curve and the HC_5. The dataset used to derive a SSD curve can be restricted and homogeneous (made up of the same effect level for the same endpoint, with the same exposure duration) or relaxed and heterogeneous (a mix of different effect levels and/or different endpoints, with varying exposure durations). However, the more restricted the data requirements are, the smaller the available dataset will be. And for most substances, with the exception of the well-tested chemicals, this may result in a dataset that is too small to create a statistically defensible SSD curve. When the SSD dataset is very homogeneous, the interpretation of the HC_5 is straightforward. If, for example, the dataset is made up of only 96-h LC_{50} data (a tempting choice, given its abundance due to the traditional preference in toxicity testing toward this endpoint), the HC_5 is the concentration where 5 % of the species in a system are expected to manifest mortality to 50 % of their population after 96 h of exposure. But it gives no information on any other harmful impacts, nor does it allow conclusions on mortality levels (other than 50 %) at other concentrations or exposure times for any species. While such a restrictive approach allows for a very explicit interpretation, it is generally not recommended when an environmentally relevant level of protection is desired. But the authoring organization has to decide whether this HC_5 would fulfill the protection goal of their jurisdiction or whether additional extrapolation (e.g., by the use of additional safety factors or change of HC value) is required.

A similar explicit interpretation applies to a homogeneous dataset of non-lethal endpoints or when a lower level of impact (i.e., <50 %) is selected.

Plotting a more heterogeneous dataset made up of various endpoints for a wider range of exposure conditions will not only create a larger dataset, and will, therefore, allow application of the SSD approach to more substances of concern, but will also better fulfill the requirement of creating a benchmark that is broadly protective. Moreover, selecting a heterogeneous dataset consisting of low-effect levels (e.g., EC_{10} and/or EC_{20}) of sensitive endpoints will yield a HC_5 that should afford a high level of protection. But the heterogeneity of the dataset no longer allows a clear interpretation of the curve, short of saying that the HC_5 is the concentration where 5 % of the species in a system are expected to manifest some (for EC_{20} data) or even no/low (for EC_{10} data) sensitive impacts to 10 or 20 % of their population.

Data aggregation

Other decisions that have to be made are how to combine multiple studies with the same endpoint for a species and (if dataset heterogeneity is allowed) how to deal with multiple endpoints for a species. In the former situation, one option is to only plot the lowest concentration; another option is to plot an averaged value (e.g., the geometric mean, or others). But in this case, consideration has to be given to the possibility that the different toxicity tests were done to differently adapted populations (e.g., one or more test populations had prior exposure and adaptation to the test substance) or to cryptic species. Averaging the values in this case is not the proper approach. In the latter issue (multiple endpoints for same species), one option is to plot only the most sensitive endpoint for a species. But because the intention of the SSD approach is to use all (or most) available data, to get a representation of the various endpoints in order to be inclusive and to derive a benchmark value that is protective with respect to all potential harmful effects, it is possible, albeit not traditional, to plot several endpoints for the same species. It will transform the species sensitivity distribution approach into an effect sensitivity distribution approach. This approach has its own issues to consider, for example, how to balance the contribution of well-studied organisms (which will have many different endpoints tested) compared to rarely studied organisms. However, the approach in itself is not wrong, and the authoring organization has to decide whether this approach is acceptable to them and fulfills their requirements.

Another consideration is to combine or split the available toxicity information according to the environmental media or ecosystems (lakes versus rivers, cold water versus warm water rivers, marine, estuarine, freshwater, etc.); according to broad taxons (plant versus animals, vertebrates versus invertebrates, etc.); or according to toxic mode of action. Splitting in this manner will allow the derivation of region-, eco-, or situation-specific benchmarks or benchmarks that address different protection goals or targets. Equally, separating out locally occurring species will allow the derivation of site-specific benchmarks. However, splitting and separating out will also reduce the size of the dataset. One aspect to consider is that the splitting or separating of one large dataset into two or more subsets often yields HC_5 values where some are lower than the HC_5 value of the combined dataset. As the desire generally is to derive a benchmark that is protective for all species under all circumstances, this poses an interesting challenge. If, for example, plotting the (site-specific) data for the plants and algae separate from the vertebrates and/or invertebrates yields a higher HC_5 for the first group and a lower HC_5 for the second group than the combined HC_5, which is the appropriate HC_5 value to use for the water quality benchmark for this site? Is the combined HC_5 more representative of the ecosystem in question than the parsed out HC_5? As not every dataset is large enough to split, when should the splitting be done? These questions need to be addressed by the authoring organization in the derivation protocol.

Statistical curve fitting

Statistics provides many different methods to fit a curve to a given dataset, each with its own strengths and weaknesses, but no method is best suited for every dataset. Some jurisdictions have decided to use one method for every dataset; other jurisdictions have decided to compare the results of several methods and choose the best-suited for a particular dataset. This necessitated identifying decision criteria and more work in the benchmark derivation process for these jurisdictions, but they have deemed that the additional effort yields better and more defensible water quality benchmarks.

Traditionally, the *Y*-intercept at the 5th percentile (the HC_5) is the value of choice for a water quality benchmark. However, this is an arbitrary choice, weighing the additional protection gained by using a lower percentile (e.g., HC_1) but increased variability and uncertainty of the curve at this level versus the lesser protection but decreased variability and uncertainty of the curve at a higher level (e.g., HC_{10} or HC_{15}).

Recognizing the impact of subjective decisions

It must be recognized that these aforementioned issues require subjective decisions that can lead to noticeable differences in final value! This important point has been investigated in an international round-robin test, with startling findings (Hahn et al. 2009, 2013). When experienced aquatic hazard assessors from around the world independently derived no-effect concentrations (i.e., equivalents to water quality benchmarks) for select substances using the same datasets, the observed variation was up to three orders of magnitude. Reasons were the obvious factors such as the size of the dataset and the methodology used, but primarily individual decisions of the developers within the scope of their respective methodology used (e.g., key study selection, acute versus chronic definitions, and size of assessment factors). Similar observations are reported by Junghans et al. (2012). This shows clearly that science can inform in these cases, but cannot make the necessary policy decisions. The proper development of a water quality benchmark does not allow blind reliance on a method, but requires a thorough understanding of the associated issues and underlying science.

Protocol summary

The authoring organization and the benchmark developers have to decide on many scientific and policy-related issues as there are many different ways to derive a water quality benchmark, all with their own benefits, problems, and issues. There is no clear best method, and a combination of several methods and approaches in a tiered manner may be the most advisable path to take as this allows flexibility to respond to different needs and opportunities.

Future opportunities

The development of water quality benchmarks requires a special expertise. It is a task which may seem simple, but it is complex in its details, and it is not easy to do it "right." The available methods have come a long way compared to their beginnings, but there is also still room for improvements.

First-generation water quality benchmarks are generally derived by multiplying the lowest (or an average of a few of the lowest) acceptable toxicity endpoint by an arbitrary safety factor. This value was deemed to estimate a benchmark protective for the aquatic environment. However, this first-generation approach does not provide any prediction of the potential environmental impacts occurring at the benchmark value, let alone at higher or lower environmental concentrations of the contaminant. Second-generation water quality benchmarks are derived by using all available and acceptable (not only the lowest) endpoints and statistical assessment tools to determine this value (species sensitivity distribution approach). The use of more data points and statistical analysis of the data is assumed to reduce the uncertainty and arbitrariness inherent in the first-generation approach. Furthermore, it now allows for a rudimentary prediction of potential effects occurring at different environmental concentrations of the contaminant. Nevertheless, none of the current approaches goes beyond providing only a single threshold value for a given environmental condition or analyzes the uncertainty around the benchmark. And even current second-generation water quality benchmarks do not provide an analysis of the risk (or even hazard) surrounding the chosen threshold concentration (i.e., the water quality benchmark) or concentrations above or below this threshold. By entering into the realm of an environmental risk assessment, a third-generation water quality benchmark may be able to provide such an analysis and allow predictions of the ecological impacts that may occur at different environmental levels of a contaminant (e.g., when a benchmark is slightly/greatly exceeded). This, in turn, would allow for an evaluation, and potentially quantification, of the societal benefits related to meeting a water quality benchmark or, similarly, the societal losses incurred when failing to maintain or achieve a water quality benchmark. However, any such increase in predictive power of a benchmark will come with an exponential increase in toxicological and ecological data requirements.

One particular area of consideration is the issue of cryptic species and adapted populations. While both issues will have a marked influence on the benchmark value, neither has been properly addressed yet in water quality benchmark derivation or environmental risk assessment. However, the inclusion of the genetic fingerprint and appropriate details on the source and origin of the test organism in the toxicity test reports is a necessary requirement for being able to incorporate them.

More and more jurisdictions develop their own water quality benchmarks. This is an indication of the need and usefulness of these benchmarks. But as it is often done in isolation, with little international cooperation, it is also a costly duplication of effort worldwide as all derivation methods follow a similar approach (data acquisition, data evaluation, benchmark derivation) and use the same toxicological data (or subsets thereof). Some jurisdictions dedicate huge resources toward benchmark development and have a large capacity to do so, while others have only a limited capacity. But the capacity to develop benchmarks is often not related to their need. There is a need for international cooperation and the transfer of expertise.

By not working together, and not transferring knowledge and results, jurisdiction duplicates in part or wholly the work already done by others. Each jurisdiction starts anew with the assembly of the available toxicity data and reevaluates the same studies already analyzed by other jurisdictions and (hopefully) ends up with the same basic set of data, augmented and expanded by the most recently published results. This process would greatly benefit from the creation of an international repository for evaluated and screened data suitable for generic and site-specific water quality benchmark derivation.

There are valid reasons for a jurisdiction to create its own derivation protocol and its own water quality benchmarks to address its specific needs. However, not all jurisdictions in need of water quality benchmarks have the capacity or expertise to develop them. It may now be time to internationally coordinate and harmonize this process. This could be done through the creation of an independent international circle of experts tasked with (a) the creation and management of an international data repository for water quality benchmark development and (b) the development of generic and ecoregion-specific water quality benchmarks for substances of international concern. Individual jurisdictions could then adopt these benchmarks, or at least use them as starting points for their derivation of benchmarks specific to their individual needs.

References

Ågerstrand M, Küster A, Bachmann J, Breitholz M, Ebert I, Rechenberg B, Rudén C (2011) Reporting and evaluation criteria as means towards a transparent use of ecotoxicity data for environmental risk assessment of pharmaceuticals. Environ Pollut 159(10):2487–2492

An Y-J, Lee J-K, Cho S (2011) Korean water quality standards for the protection of human health and aquatic life. http://www.wepa-db.net/pdf/0712forum/paper10.pdf. Accessed 27 May 2013

ANZECC/ARMCANZ (Australian and New Zealand Environment and Conservation Council and Agriculture and Resource Management Council of Australia and New Zealand) (2000) Australian and New Zealand guidelines for fresh and marine water quality. Canberra, Australia

Australian Government (2013) National water quality management strategy. http://www.environment.gov.au/water/policy-programs/nwqms/index.html (web site updated 30 April 2013)

Barata C, Agra AR, Soares AMVM (2012) Does genetic adaptation matter?—a hypothesis tested using life-history consequences of adaptation and acclimatization to copper of *Daphnia longispina*. Poster at SETAC World Conference, Berlin, Germany, 2012. cbmqam@cid.csic.es

CCME (Canadian Council of Ministers of the Environment) (1991) Appendix IX—a protocol for the derivation of water quality guidelines for the protection of aquatic life (April 1991). In: Canadian water quality guidelines. Canadian Council of Resource and Environment Ministers, 1987. Prepared by the Task Force on Water Quality Guidelines (updated and reprinted with minor revisions and editorial changes in Canadian environmental quality guidelines, chapter 4. Canadian Council of Ministers of the Environment, 1999, Winnipeg), 10 pp

CCME (Canadian Council of Ministers of the Environment) (1999) Canadian environmental quality guidelines. Canadian Council of Ministers of the Environment, Winnipeg

CCME (Canadian Council of Ministers of the Environment) (2003) Canadian water quality guidelines for the protection of aquatic life: guidance on the site-specific application of water quality guidelines in Canada: procedures for deriving numerical water quality objectives. In: Canadian environmental quality guidelines, 1999. Canadian Council of Ministers of the Environment, 1999, Winnipeg, 146 pp

CCME (Canadian Council of Ministers of the Environment) (2007a) A protocol for the derivation of water quality guidelines for the protection of aquatic life 2007. In: Canadian environmental quality guidelines, 1999. Canadian Council of Ministers of the Environment, 1999, Winnipeg, 37 pp

CCME (Canadian Council of Ministers of the Environment) (2007b) Strategic planning guidance—Water Quality Task Group, 2007. Canadian Council of Ministers of the Environment, Winnipeg, 42 pp

CCME (Canadian Council of Ministers of the Environment) (2013a) Organisational chart on website. http://www.ccme.ca/assets/pdf/orgchart.pdf (updated 1 April 2013)

CCME (Canadian Council of Ministers of the Environment) (2013b) Water quality index. http://www.ccme.ca/ourwork/water.html?category_id=102 (updated 29 August 2013)

CCREM (Canadian Council of Resource and Environment Ministers) (1987) Canadian water quality guidelines. Winnipeg

Chapman PF, Crane M, Wiles JA, Noppert F, McIndoe EC (eds) (1995) Asking the right questions: ecotoxicology and statistics. The report of a Workshop held at Royal Halloway University of London, Egham, Surrey, United Kingdom, 26–27 April 1995. Society of Environmental Toxicology and Chemistry—Europe, Brussels

Chapman PM, Caldwell RS, Chapman PF (1996) A warning: NOECs are inappropriate for regulatory use. Environ Toxicol Chem 15:77–79

Chapman PM, Fairbrother A, Brown D (1998) A critical evaluation of safety (uncertainty) factors for ecological risk assessment. Environ Toxicol Chem 17:99–108

Dourson M (2005) Uncertainty factors. In: Encyclopedia of toxicology (second edition). Elsevier Inc., Amsterdam, p. 401–406. ISBN: 978-0-12-369400-3

Duan Y, Guttman SI, Oris JT (1997) Genetic differentiation among laboratory populations of *Hyalella azteca*: implications for toxicology. Environ Toxicol Chem 16:691–695

EC (Environment Canada) (2013) Federal environmental quality guidelines—alcohol ethoxylates. Environment Canada, Ottawa, http://www.ec.gc.ca/ese-ees/default.asp?lang=En&n=164786DB-1

EC (European Communities) (2000) Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. OJ L327/1, 22.12.2000

EC (European Communities) (2011) Technical guidance for deriving environmental quality standards—common implementation strategy for the Water Framework Directive (2000/60/EC). Guidance document No. 27, Technical Report 2011-055, European Communities, Brussels, Belgium. https://circabc.europa.eu/w/browse/a3c92123-1013-47ff-b832-16e1caaafc9a. Accessed 01 March 2013

ECHA (European Chemicals Agency) (2012) Practical guide 3: how to report robust study summaries, version 2.0. ECHA-10-B-06.1-EN, European Chemicals Agency. http://echa.europa.eu/. Accessed 24 May 2013

Elmegaard N, Jagers op Akkerhuis GJAM (2000) Safety factors in pesticide risk assessment. Differences in species sensitivity and acute–chronic relations. NERI Technical Report No. 325. National Environmental Research Institute, Silkeborg, Denmark, 60 pp

Environment Agency (2007) Prioritising chemicals for standard derivation under Annex VIII of the Water Framework Directive—Science Report SC040038/SR. Environment Agency, Bristol, BS32 4UD, UK, 152 pp

Feckler A, Thielsch A, Schwenk K, Schulz R, Bundschuh M (2012) Differences in the sensitivity among cryptic lineages of the *Gammarus fossarum* complex. Sci Total Environ 439:158–164

Feng CL, Wu FC, Zhao XL, Li HX, Chang H (2012) Water quality criteria research and progress. Sci China Earth Sci 55(6):882–891

Gabaldón C, Montero-Pau J, Serra M, Carmona MJ (2013) Morphological similarity and ecological overlap in two rotifer species. PLoS ONE 8(2):e57087. doi:10.1371/journal.pone.0057087

Hahn T, Stauber J, Dobson S, Howe P, Kielhorn J, Koennecker G, Diamond J, Lee-Steere C, Schneider U, Sugaya Y, Taylor K, Van Dam R, Mangelsdorf I (2009) Reducing uncertainty in environmental risk assessment (ERA): clearly defining acute and chronic toxicity tests. Integr Environ Assess Manag 3:175–177

Hahn T, Diamond J, Dobson S, Howe P, Kielhorn J, Koennecker G, Lee-Steere C, Mangelsdorf I, Schneider U, Sugaya Y, Taylor K, van Dam R, Stauber J (2013) Predicted no effect concentration (PNEC) derivation as a significant source of variability in environmental hazard assessments of chemicals in aquatic systems: an international analysis. Integr Environ Assess Manag. doi:10.1002/ieam.1473. Accepted 3 August 2013

Hoekstra JA, Van Ewijk PH (1993) Alternatives for the no-observed effect level. Environ Toxicol Chem 12:187–194

Hogg ID, Larose C, de Lafontaine Y, Doe KG (1998) Genetic evidence for a *Hyalella* species complex within the Great Lakes–St. Lawrence River drainage basin: implications for ecotoxicology and conservation biology. Can J Zool 76:1134–1140

Junghans M, Von Arb S, Whitehouse P, Johnson I (2012) Variability in environmental quality standards—how much is there and what are the causes? Poster at SETAC World Conference, Berlin, Germany, 2012. Marion.junghans@oekotoxzentrum.ch or paul.whitehouse@environment-agency.gov.uk

Klimisch HJ, Andreae M, Tillman U (1997) A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul Toxicol Pharmacol 25:1–5

Lepper P (2005) Manual on the methodological framework to derive environmental quality standards for priority substances in accordance with Article 16 of the Water Framework Directive (2000/60/EC). Fraunhofer Institute Molecular Biology and Applied Ecology, Schmallenberg, Germany, 15 September 2005

Malkiewicz K, Hansson SO, Rudén C (2009) Assessment factors for extrapolation from short-time to chronic exposure—are the REACH guidelines adequate? Toxicol Lett 190:16

Mensink BJWG, Smit CE, Montforts MHMM (2008) Manual for summarising and evaluating environmental aspects of plant protection products. RIVM report no. 601712004/2008. RIVM, Bilthoven, the Netherlands. www.rivm.nl. Accessed 27 May 2013

Moore DRJ, Caux P-Y (1997) Estimating low toxic effects. Environ Toxicol Chem 16(4):794–801

Nys C, Janssen CR, De Schamphelaere K (2012) A comparison of the chronic Pb toxicity between laboratory and field populations of the great pond snail (*Lymnaea stagnalis*). Poster at SETAC World Conference, Berlin, Germany, 2012. Chnys.nys@ugent.be

OECD (Organisation for Economic Cooperation and Development) (2005) Guidance document on the validation and international acceptance of new or updated test methods for hazard assessment. OECD, Paris

Pfenninger M, Schwenk K (2007) Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. BMC Evol Biol 7:121. doi:10.1186/1471-2148-7-121

Pieters MN, Kramer HJ, Slob W (1998) Evaluation of the uncertainty factor for subchronic-to-chronic extrapolation: statistical analysis of toxicity data. Regul Toxicol Pharmacol 27(2):108–111

Pohl HR, Chou C-HSJ, Ruiz P, Holler JS (2010) Chemical risk assessment and uncertainty associated with extrapolation across exposure duration. Regul Toxicol Pharmacol 57(1):18–23

Posthuma L, Suter II GW, Traas T (eds) (2002) Species sensitivity distribution in ecotoxicology. CRC, Boca Raton, 587 pp

Rand GM (1995) Fundamentals of aquatic toxicology: effects, environmental fate, and risk assessment, 2nd edn. Taylor and Francis, Washington, 1125 pp

Renwick AG (1995) The use of an additional safety or uncertainty factor for nature of toxicity in the estimation of acceptable daily intake and tolerable daily intake values. Regul Toxicol Pharmacol 22(3):250–261

RIVM (Reijksinstituut voor Volksgezondheid en Milieu) (2011) Evaluatie van de methodiek voor het afleiden van indicatieve milieurisicogrenzen—RIVM Rapport 601357006/2011. Ministerie van Volksgezondheid, Welzijm en Sport, p 68

Rocha-Olivares A, Fleeger JW, Foltz DW (2004) Differential tolerance among cryptic species: a potential cause of pollutant-related reductions in genetic diversity. Environ Toxicol Chem 23(9):2132–2137

Stephan CE, Mount DI, Hansen DJ, Gentile JH, Chapman GA, Brungs WA (1985) Guidelines for deriving numerical national water quality criteria for the protection of aquatic organisms and their uses. United States Environmental Protection Agency, EPA-822-R85100, Washington, DC

US EPA (United States of America Environmental Protection Agency) (2013) Antidegradation policy. http://water.epa.gov/scitech/swguidance/standards/adeg.cfm (updated 6 March 2012)

Van Herwijnen R, Postma J, Keijzers R, Van Leeuwen L (2012) Comparison of environmental quality standard derivation methods: indicative versus WFD methodology. SETAC Poster TH282. Rene.van.herwijnen@rivm.nl

Warne M, Stauber J, Van Dam R (2008) NOEC and LOEC data should no longer be generated or used. Australas J Ecotoxicol 14:1–5

WEPA (Water Environment Partnership in Asia) (2012) Outlook on water environmental management in Asia 2012. Published by the Ministry of the Environment, Japan, Tokyo. ISBN: 978-4-88788-108-2. http://www.wepa-db.net/pdf/1203outlook/01.pdf

Wu F, Meng W, Zhao X, Li H, Zhang R, Cao Y, Liao H (2010) China embarking on development of its own national water quality criteria system. Environ Sci Technol 44(21):7992–7993

Yamazaki K (2011) Regulatory standards for conservation of aquatic life in Japan. Environmental Health Department, Ministry of the Environment, Japan, Tokyo, Japan. Keynote Lecture at EQSPAE-2011. International Conference on Deriving Environmental Quality Standards for the Protection of Aquatic Ecosystems, 3–7 December 2011, University of Hong Kong. Kunihiko_yamazaki@env.go.jp