Environ Sci Pollut Res (2014) 21:6–16
DOI 10.1007/s11356-013-1732-8

ENVIRONMENTAL QUALITY BENCHMARKS FOR PROTECTING AQUATIC ECOSYSTEMS

Water quality guidelines for chemicals: learning lessons to deliver meaningful environmental metrics

Graham Merrington · Youn-Joo An · Eric P. M. Grist · Seung-Woo Jeong · Chuthamat Rattikansukha · Susan Roe · Uwe Schneider · Suthipong Sthiannopkao · Glenn W. Suter II · Rick Van Dam · Patrick Van Sprang · Ju-Ying Wang · Michael St. J. Warne · Paul T. Yillia · Xiao-Wei Zhang · Kenneth M. Y. Leung

Received: 6 January 2013 / Accepted: 9 April 2013 / Published online: 26 April 2013

Abstract Many jurisdictions around the globe have well-developed regulatory frameworks for the derivation and implementation of water quality guidelines (WQGs) or their equivalent (e.g. environmental quality standards, criteria, objectives or limits). However, a great many more still do not have such frameworks and are looking to introduce practical methods to manage chemical exposures in aquatic ecosystems. There is a potential opportunity for learning and sharing of data and information between experts from different jurisdictions in order to deliver efficient and effective methods to manage potential aquatic risks, including the considerable reduction in the need for aquatic toxicity testing and the rapid identification of common challenges. This paper reports the outputs of an international workshop with

Responsible editor: Hailong Wang

G. Merrington (✉) · E. P. M. Grist
WCA Environment Limited, Brunel House, Volunteer Way, Faringdon, Oxfordshire SN7 7YR, UK
e-mail: graham.merrington@wca-environment.com

Y.-J. An
Department of Environmental Science, Konkuk University, 1 Hwayang Dong, Seoul 143701, South Korea

S.-W. Jeong
Department of Environmental Engineering, Kunsan National University, Kunsan 573701, South Korea

C. Rattikansukha
Pollution Control Department, Ministry of Natural Resource and Environment, Bangkok, Thailand

S. Roe
National Guidelines and Standards Office, Environment Canada, Gatineau, QC, Canada

U. Schneider
Independent Environmental Consultant, 1944 Forced Road, Ottawa K0A3H0 ON, Canada

S. Sthiannopkao
Department of Environmental and Occupational health, National Cheng Kung University, Tainan City, Taiwan

G. W. Suter II
National Centre for Environmental Assessment, U.S. Environmental Protection Agency, Cincinnati, OH, USA

R. Van Dam
Environmental Research Institute of the Supervising Scientist, GPO Box 461, Darwin, Australia

P. Van Sprang
ARCHE, Stapelplein 70 box 104, 9000 Ghent, Belgium

J.-Y. Wang
National Marine Environmental Monitoring Center, State Oceanic Administration, 42 Linghe Street, Dalian 116023, China

M. St. J. Warne
Water Quality and Investigations, Environmental Monitoring and Assessment Science, Science Delivery, Department of Science, Information Technology, Innovation and the Arts, GPO Box 5078, Brisbane, QLD 4001, Australia

P. T. Yillia
Institute of Water Quality, Vienna University of Technology, Karlspl 13-226, A-1040 Vienna, Austria

X.-W. Zhang
School of the Environment, Nanjing University, 22 Hankou Road, Nanjing, Jiangsu 210023, China

K. M. Y. Leung (✉)
The Swire Institute of Marine Science and School of Biological Sciences, The University of Hong Kong, Pokfulam, Hong Kong, China
e-mail: kmyleung@hku.hk

representatives from 14 countries held in Hong Kong in December 2011. The aim of the workshop and this paper was to identify 'good practice' in the development of WQGs to deliver to a range of environmental management goals. However, it is important to broaden this consideration to cover often overlooked facets of implementable WQGs, such as demonstrable field validation (i.e. does the WQG protect what it is supposed to?), fit for purpose of monitoring frameworks (often an on-going cost) and finally how are these monitoring data used to support management decisions in a manner that is transparent and understandable to stakeholders. It is clear that regulators and the regulated community have numerous pressures and constraints on their resources. Therefore, the final section of this paper addresses potential areas of collaboration and harmonisation. Such approaches could deliver a consistent foundation from which to assess potential chemical aquatic risks, including, for example, the adoption of bioavailability-based approaches for metals, whilst reducing administrative and technical burdens in jurisdictions.

Keywords Water quality guidelines · International collaboration · Harmonisation · Water quality management · Environmental quality standards

Introduction

This synthesis paper is aimed at identifying 'good practice' in the development of water quality guidelines (WQGs). Furthermore, it details critical aspects which are often overlooked during the implementation WQGs. In this paper, WQGs, water quality criteria, water quality objectives, environmental quality standards and other such titles are used to encompass numerical limits of a particular pollutant or group of pollutants in water, sediment or biota which should not be exceeded in order to protect human health and aquatic ecosystems.

Some jurisdictions around the world such as North America, Europe, Australia and New Zealand have mature, well-developed regulatory frameworks for the management of chemicals in the aquatic environment. Other jurisdictions like China, Hong Kong, Korea, Japan and South Africa have recently embarked on developing their own WQG system (An et al. 2008; Jin et al. 2009; Wu et al. 2010; Wepener and Chapman 2012). Whilst specific details of these frameworks differ between jurisdictions, key elements are generally conserved, WQGs are derived and monitoring in the environment is then undertaken to ensure that guidelines are not exceeded, or identify instances where regulatory intervention are required (e.g. Stephan et al. 1985; US EPA 1985; ANZECC and ARMCANZ 2000a; CCME 2007; EC 2011). These jurisdictions not only have extensive experience in WQG derivation, but often the processes of derivation and implementation have been refined over many years based on practical experience (e.g. CCME 2007). This process has often involved research into the various methods of setting and implementing WQGs (Environment Agency 2009). Importantly, not all these methods have been successful due to a number of scientific and technical reasons (Environment Agency 2008). These experiences and lessons can be beneficial to other less experienced jurisdictions.

There is potential for collaboration and sharing best practice among jurisdictions. A clear example of such collaboration would be the use of common, quality-rated ecotoxicity data (e.g. OECD's Safety Information Data Sheets). In addition, there remain considerable challenges common to both mature and developing WQG jurisdictions, such as setting limit values for poorly water-soluble chemicals (and associated monitoring in biota), handling different responses to chemical exposure observed between temperate and tropical aquatic species and between native and non-native species, as well as issues associated with mixture toxicity or the management of breakdown/degradation products.

To address these issues, a special workshop was organized during the first International Conference on Deriving Environmental Quality Standards for the Protection of Aquatic Ecosystems (EQSPAE-2011) which was held in December 2011 in Hong Kong. It was attended by over 150 representatives from academia, industry and government from 14 different countries. The objectives of the workshop were to identify best practice methods for the development of WQGs so as to deliver a range of environmental management goals, as well as to identify and discuss associated activities required for successful WQG derivation and implementation, including:

- Validation of the chemical WQGs—How might this be undertaken and what data would be needed in order to fulfil scientific and regulatory requirements?
- Monitoring requirements—What to monitor, to what analytical limit of detection? Where? When and for what purpose?
- Compliance assessment—What do we do with the WQGs and the monitoring data once we have created them?
- Harmonisation—How can the resources and expertise invested in the derivation process and the lessons learnt in implementation of WQG be shared?

Deriving WQGs for ecosystem protection

When comparing and appraising the different methodologies used to derive WQGs, a key consideration is the ultimate protection goal of the WQG as these often differ greatly between jurisdictions and can therefore reflect genuine differences in methodological paradigm required to derive a WQG. Arguably, the choice of protection objective is up to the policy makers of a given jurisdiction (e.g. Directive 2000/60/EC). However, once the protection objective is set, the method for deriving a guideline should reflect this. For instance, a management goal to protect aquatic organism populations would likely require different data inputs and methodology to derive than a management goal to protect aquatic community function. It is often believed that the goal should be protection of aquatic ecosystem services, although this goal lacks a clear and explicit definition. However, the environmental laws in many nations require protection of aquatic ecosystems, irrespective of their services to humans (e.g. Directive 2000/60/EC).

A great deal of guidance already exists on the derivation of WQGs from many jurisdictions (e.g. CCME 2007; EC 2011). Many similarities exist between the various methodologies used for the identification, collation and quality assessment of relevant ecotoxicity data (e.g. CCME 2007; ANZECC and ARMCANZ 2000a) and those used for the assessments of residual uncertainty (e.g. CCME 2007; EC 2011).

What is very clear is that comparisons between methods should not become hampered by terminology and nomenclature. Terms such as 'standard', 'criterion', 'objective', 'limit value' and 'guideline' do not have unique globally accepted definitions and will be defined in different jurisdictions by legal terminology and in relation to management goals. Often different terminology will have consistent meaning across jurisdictions. Given this, there is clearly a need to have statements on the meaning of the terms and the context in which they have been derived (Ministry of the Environment, New Zealand 2003).

At present, probabilistic approaches for the treatment of ecotoxicity data, such as species sensitivity distributions (SSDs), are generally the preferred tool for regulators to derive WQGs rather than deterministic approaches using assessment or safety factors to account for residual uncertainty. SSDs provide a relatively simple and understandable means of converting the large body of species-level ecotoxicity data into a single community metric (i.e. the hazardous concentrations 5 %, HC5), which corresponds to a concentration at which 5 % of species would be expected to be adversely affected after exposure to a chemical (Posthuma et al. 2001; Wheeler et al. 2002). However, SSD methodologies tend to only deal with interspecies differences, and they make a number of assumptions. A summary of the assumptions made and the potential implications of these on the results of SSDs is presented in Warne (1998) and the references therein.

Though it is possible to apply an SSD approach with relatively small datasets, they are generally only considered to generate robust summary statistics when minimum criteria for the number of species and taxonomic groups are met, which vary between jurisdictions. The criteria for a robust SSD approach used in the EU and the USA are based on the conclusions of the London Workshop, an expert working group that met in 2001 (London Workshop 2001). Where it is not acceptable to apply an SSD approach in a particular jurisdiction, normally due to the lack of data from an appropriate number of species and/or taxonomic groups, jurisdictions generally fall back on applying the use of assessment/safety factors (generally ranging from 10 to 10,000) on the most conservative ecotoxicity data for a chemical. This approach offers benefits when used in an iterative risk assessment process as it is simple and transparent, and permits expert judgement with an aim of protecting all species. However, the assessment factor approach can present difficulties if used for deriving limits that are to be enshrined in legislation because the resulting guidelines tend to be relatively precautionary when compared to probabilistic approaches and dependent on the size of the assessment/safety factor applied. As such, there have been recent discussions in Europe as to the use of weight-of-evidence approaches, including field and higher tier data, and the need to have larger minimum ecotoxicity datasets for deriving WQGs (SCHER 2010).

Whilst SSDs require larger amounts of ecotoxicity data than the deterministic methods, there is a strong regulatory desire, across many jurisdictions, to reduce the amount of ecotoxicity testing, especially for vertebrates due to ethical reasons. One way to reconcile these contradictory aspirations is to extract more ecologically relevant information from SSDs, including: better use of the confidence limits (FMWQ Working Group 4. 2010), weighting of ecological keystone species (Forbes and Forbes 1993) or the relevance or prevalence of species (Hose 2005), and adopting approaches where the entire concentration–response curve and its associated uncertainty can be captured in the analysis and ultimate derivation of a WQG (e.g. hierarchical Bayesian approaches; see Gronewold and Borsuk (2010) for example).

In parts of the world, it is now recognised that the most relevant approach is to define no-effects concentrations (NECs; Fox 2010). As long ago as 1998 (Warne 1998), the proposed method for deriving the Australian and New Zealand WQGs highlighted the limitations of NOEC data and that their use in deriving WQGs be replaced by ECx type data as such data become available. Subsequently, there was strong agreement within the Working Group Revising the Australian and New Zealand WQGs for toxicants and sediments (FMWQ Working Group 4. 2010) to continue moving away from using NOECs, to EC10 or IC10 in the short term, but in the longer term to change to using NECs, providing methods of sufficient rigour, robustness, accuracy and ease of use can be developed. At the moment, a combination of ECx type data with NOECs is recognised as the best transitional option as these tend to be the most common forms in which chronic ecotoxicity data are expressed. This move away from NOECs is consistent with the many authors who have criticised their use since the 1980s with some and even calling for their use to be banned (e.g. Newman 2008; Warne and Van Dam 2008; Landis and Chapman 2011; Jager 2012).

In moving to ECx type and NEC data, there may be the need to modify the experimental design of ecotoxicity tests, with more treatments at the bottom end of concentration–response relationships. The extra effort required by these changes can be offset by reducing the degree of replication (Stephan and Rogers 1985; Moore and Caux 1997), although some indication of an adequate level of repeatability is still desirable. For pre-existing data, the ultimate preferred order of estimates for Australia and New Zealand is recommended to be EC10>BEC10>EC15-20> unconverted NOECs>converted NOECs from maximum acceptable toxic concentrations [MATC=geometric mean of NOEC and lowest observable effect concentration (LOEC)], LOECs or median lethal concentrations (LC50s) (FMWQ Working Group 4. 2010). It was also agreed that NOECs would be discarded once sufficient NEC or EC10 values ($n \geq 8$) were available for use in SSDs, but the effect of this might need to be examined if the original dataset was much larger and more representative.

Field-based and functionality-based SSDs may be used to derive WQGs or to provide lines of evidence in weight-of-evidence-based approaches (Peters et al. 2010; Peters and Simpson 2012). Field-based methods have been developed in the USA and Great Britain (Linton et al. 2007; Peters et al. 2010, 2011), but they are new and have not been widely adopted and perhaps should play a greater role in WQG derivation or validation. In Australia and New Zealand (ANZECC/ARMCANZ 2000a, b), unlike in other jurisdictions the highest quality WQGs are considered to be those derived directly using field, meso- or microcosm-derived data, although, as yet, there are no such instances of guidelines being derived on this basis. Yet in Alberta (Chambers et al. 2006) and the UK (Peters et al. 2011) guidelines have been derived in this way. However, field, meso- or microcosm-derived data have been used to determine if the WQGs derived using laboratory-based data provide adequate protection (ANZECC and ARMCANZ 2000a). This is a similar process by which higher-tier data are used in Europe.

A pragmatic suggestion to efficiently use data is to adopt tiered WQGs with associated levels of known uncertainty or reliability. For example, for screening assessments less reliable (but conservative) WQGs could be used, whereas for compliance assessments potentially leading to prosecution or other regulatory intervention, robust, relatively definitive WQGs are needed. The Australian and New Zealand WQGs (ANZECC and ARMCANZ 2000a, b) adopt a screening approach, and where a guideline trigger value is exceeded, further studies will be undertaken to identify the biological impact. This may include chemical measurements of bioavailability, modelling or ecotoxicity testing. For such screening purposes, three levels of WQGs are derived with high, moderate and low reliability based on the type (acute/chronic), quantity and taxonomic representativeness of the data (i.e. no. of species and taxa at the phyla level). This concept, while helpful, needs to more adequately address the reliability of the WQGs and will be developed further as part of the current review of the Australian and New Zealand WQGs (FMWQ Working Group 4 2010). The Canadian WQGs now also have three categories (previously two), based on data quality and quantity and preferentially employing an SSD approach. However, the Canadian and the Australian and New Zealand guidelines are not enshrined in legislation, perhaps providing flexibility not available to other jurisdictions.

Finally, a clear need for guidance has been identified on the use of non-native species (and implicitly species from other climatic zones) when deriving WQGs. SSDs are models of the distribution of sensitivities of taxa (i.e. an assembly of tested species), not descriptions of a particular community. However, regulatory concerns have led to the exclusion of non-native test species which could perhaps serve as surrogates for native species for which data are lacking and help to define the distribution (e.g. Alberta Environmental Protection 1996). Numerous studies have investigated this issue, but most have had limited applicability due to either the small number of species or chemicals used in the comparisons. The larger studies have indicated that there are differences in the sensitivity of species, but it appears to be species and chemical specific (e.g. Hobbs et al. 2004; Chapman et al. 2006: Kwok et al. 2007). Chapman et al. (2006) concluded that sensitivity data from one region will not be universally protective of other regions. A synthesis of studies that have looked at this issue would be useful, i.e. are there geographic patterns of sensitivity or is it possible to generalise on some other basis?

Validation of water quality guidelines

It is clear that many existing WQGs, derived exclusively from laboratory ecotoxicity data, have not been assessed under field conditions to establish whether they are indeed protective of the ecosystems which they are intended to protect and not overly conservative. This benchmarking of the WQG against the biological response in the field is what is considered here to mean ‘validation’.

Other than for the data-rich metals (e.g. Cu, Ni) and some historic, but extensively monitored biocides (e.g. tributyltin) and pesticides (e.g. cypermethrin), the data needed to demonstrate the protectiveness of WQG in the field are seldom available. These validation data could come from microcosm and mesocosm studies or from multi-species, field-based studies. In Europe, mesocosm data are used in a weight-of-evidence approach on decisions in relation to the selection of appropriate assessment factors (EC 2012). The Environment Agency of England and Wales has also used field-based data to derive a WQG for iron, for as well as having both a physical and chemical effect the associated laboratory ecotoxicity data are of limited quality (Peters et al. 2011). In addition, mathematical (focussing on population-level and community-level effects), direct toxicity measurements and ecosystem models (e.g. food web models) could help to explain the relationship between data derived from the laboratory and field (Liney et al. 2006). All have potential drawbacks, but used in weight-of-evidence approaches or as part of a 'tool box' of methods and approaches, they could deliver increased understanding of the links between chemical and biological metrics to reduce the uncertainty associated with management decisions. Also, there is a need to account for additional pressures and confounding factors when considering field data (e.g. through the use of limiting functions, such as piece-wise or quantile regression; Crane et al. 2007). The derivation, validation and acceptance of WQGs are further complicated through the conundrum that WQGs are generally derived using laboratory toxicity tests by exposing a species or a group of species to a single chemical stressor, while in reality, various species in a community are simultaneously exposed to multiple toxicants and stressors (although often field-based data will reflect multiple exposures and stressors; Leung et al. 2005; Peters et al. 2011).

It is likely that for some jurisdictions, biological data, appropriate for validation exercises, have been collected in order to fulfil other regulatory requirements. For example, in Europe, the Water Framework Directive requires Member States to use several well-developed classification tools to assess ecological status of aquatic habitats (e.g. benthic invertebrates, macrophytes, diatoms, fish, etc.), and these may be appropriate for WQG validation exercises. There are several examples of using these types of data to validate WQGs (Crane et al. 2007; Linton et al. 2007; Peters et al. 2010). The applicability of such tools and approaches outside Europe and the USA, however, requires further investigation.

In addition to the validation of WQGs, there is also a need to validate those tools and models used to aid the interpretation of monitoring data or to increase the relevance of WQGs, such as those that account for bioavailability. For metals, this generally means biotic ligand models (BLMs) or empirical or operationally defined relationships, such as those that account for the mitigating effects of dissolved organic carbon (e.g. the proposed environmental quality standard of lead for the Water Framework Directive, EC 2012) or water hardness (USEPA 1985; ANZECC and ARMCANZ 2000a, b). Indeed, BLMs have largely been developed and validated for application to metals in freshwaters in Europe and North America, but with most applications to acute effects and only in limited cases to chronic effects. Applicability outside these regions, e.g. in subtropical and tropical waters, will therefore require additional validation to ensure that predictions remain within acceptable limits. Research projects to validate the Ni BLM in Australia (Chris Schlekat, NiPERA, personal communication) and the Cu BLM in China (Z.T. Liu, Chinese Academy of Environmental Science, personal communication) are currently being conducted.

If similar bioavailability models exist for organic chemicals in waters, there is a need to demonstrate that the applicability of the models is suitable outside the original development range. Finally, it is clear that whatever approach is taken to validate a WQG, a clear set of unambiguous success criteria for the validation should be established prior to the exercise being undertaken. For example, is it enough to demonstrate that the WQG is protective in the field for the most sensitive species identified from the laboratory data?

Considerations in establishing and running monitoring programs

A great deal of the focus often falls upon the scientific derivation of a WQG, but the utility of WQGs in supporting management decisions is equally a function of the quality and quantity of the monitoring data to which it is compared. The clear aims and data quality objectives (e.g. US EPA 1994) of the environmental monitoring program need to be determined at an early stage, and perhaps, the experience of jurisdictions with mature monitoring programs could be useful in establishing best practices. Furthermore, the monitoring statistic associated with the WQG (e.g. annual average) must match the monitoring regime. Where monitoring data will be compared to a WQG, an exceedence (e.g. any single value above or 20 % of monthly averages more than 10 % above) and its consequence (e.g. does an exceedence trigger management action, further investigation?) should be defined (Crane et al. 2010).

Whilst biological and chemical sampling is to be undertaken as part of a monitoring program, it is useful if samples are collected as close together as possible, both in space and time. This may seem odd, but historically, this has not always been the case as locations for chemical and ecological sampling are not always mutually compatible (e.g. Environment Agency of England and Wales; Graham Merrington, WCA, personal communication). Further, supporting (and relatively cheap) abiotic parameters may also be determined, such as pH, dissolved organic carbon (DOC), alkalinity, hardness, etc., in order to facilitate interpretation of contaminant data (especially for metals) in terms of bioavailability. If data are lacking, provisional screening exercises may be undertaken using surrogates (e.g. DOC predictions from dissolved iron; Peters et al. 2011), other site data (e.g. FOREGS for EU waters) and defaults based on historical data for the same locations (Environment Agency 2009).

Sampling location is often defined on the basis of "purposive selection", i.e. sampling is undertaken where an issue may be expected. Effectively, these 'hazard-based' assessments can be expanded to ensure effective use of resources. Often this is realized using tiered approaches, i.e. starting from relatively simple but ecologically less relevant, towards more complex but ecologically more relevant methods. For example, for metals, precautionary monitoring programmes may initially sample 'total' trace element concentrations, undertaking dissolved measurements as a first refinement step and only initiating bioavailability measurement, modelling and/or direct toxicity assessments at sampling sites where potential risks are identified (ANZECC and ARMCANZ 2000a; Fig. 1).

A pragmatic approach to account for ambient background concentrations (ABCs) is the so-called added risk approach (Crommentuijn et al. 1997). Ideally, local sampling sites beyond the immediate area of anthropogenic or geogenic enrichment need to be sampled to provide data from which estimates of ABCs can be made (International Organisation for Standardisation 2005a, b). Other possible methods for the estimation of ABCs could also be useful, such as trace metal concentrations in deep groundwater, or the use of databases of concentrations of chemicals in waters from relatively pristine sites (e.g. FOREGS database). Alternately, modelling methods used to estimate ABCs in soils (Hamon et al. 2004; Zhao et al. 2007) could, with modification, offer potential. The estimation of ABCs is, arguably, a policy-based approach to deliver a pragmatic fix for local assessments and management decisions. A tiered approach for metals which accounts for both bioavailability and ABCs has been developed under the EU WFD (Van Sprang et al. 2008) and for soil quality guidelines in Australia (NEPC 2011). Moreover, it is also important to verify if organisms that are naturally exposed to elevated concentrations of chemical contaminants have developed tolerance.

Having defined the purpose of monitoring, it is imperative to establish the sampling frequency. This should include accounting for both seasonal factors such as rainfall, but also the usage pattern of the contaminant being monitored. For example, if a plant protection production (i.e. using

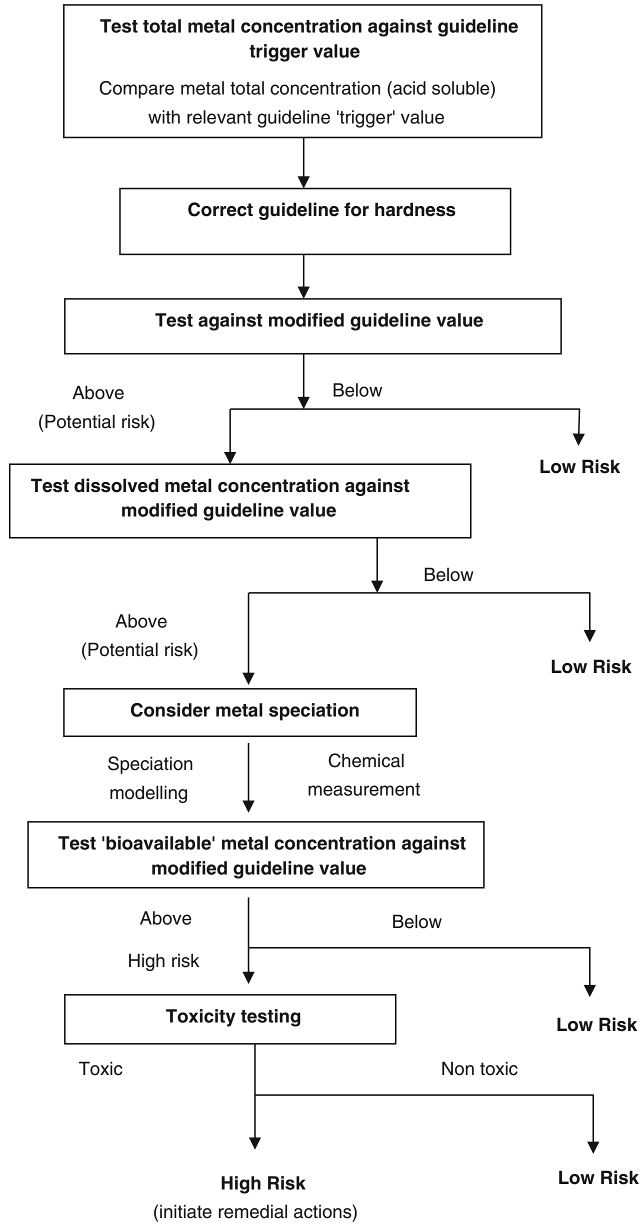

Fig. 1 Hierarchical decision tree for metals in surface waters (modified from ANZECC and ARMCANZ 2000a)

pesticides) is only conducted at certain periods in the cropping cycle, then monitoring should attempt to capture these peak potential exposure periods. The frequency of monitoring and how these monitoring data should be used are discussed in the next section.

Assessment: the use of monitoring data and WQGs to support decisions on water quality management

Irrespective of the maturity of the regulatory regime, there are some clear issues that arise when assessing whether a

WQG has been exceeded. Collecting and analysing water samples is a relatively expensive process. However, if some of these data are reported as less than the limit of detection or below the minimum reporting value, then it can be difficult to derive a meaningful summary statistic for the environmental concentration. Commonly, a value of 1/2 the minimum reporting value is substituted for these values when calculating the mean or other summary statistics from such data. However, this leads to substantial bias on any subsequently calculated statistical values (Singh and Nocerino 2002) and represents a significant loss of information (Helsel and Cohn 1988). Despite the clear limitations of such data substitution, it is specified in the European QA/QC Directive (EC 2009) as an appropriate data manipulation. An alternative to substitution is to use a statistical technique to calculate a dataset's descriptive statistics (i.e. mean, standard deviation and percentiles) incorporating censored data (Helsel 2005). In medical and industrial statistics, the standard method for calculating descriptive statistics from censored data is the Kaplan–Meier (KM) method (Klein and Moeschberger 2003a, b; Meeker and Escobar 1998). This is a non-parametric method designed to incorporate data with multiple censoring levels and does not require an assumed data distribution (e.g. log-normal), unlike similar maximum likelihood (MLE) techniques. KM estimates percentiles (including the median) and the mean of the complete dataset, including the censored data, from a cumulative distribution function. Despite the availability of these methods, Newman (1995) considers that the use of datasets comprised of ≥40 % censored data as unreliable and possibly meaningless.

"Face-value" compliance assessment is where the average concentration of a chemical in grab samples taken over 1 year is compared directly with the WQG. In Europe, the annual average is calculated as the arithmetic mean of 12 monthly samples taken over the period of 1 year. However, such a comparison is relatively crude in that it does not provide any statistical confidence in assigning a failure or exceedance of the WQG. Some Member States in the EU address this by accounting for the variation in the monitoring data (i.e. the standard deviation of the data) to provide a measure of the statistical confidence of failure (e.g. ISO/WD 5667–20. 2005). This is especially important when critical management decisions are to be made based on a WQG exceedance (i.e. only exceedances of >95 % confidence result in regulatory intervention).

In addition, monitoring data are not equal in terms of quality. While schemes exist to assess the quality and reliability of effects data, there are fewer schemes that are available for the assessment of monitoring data. An example of one of these schemes is reported by ANZECC and ARMCANZ (2000b) which gives guidance on quality assurance and quality control protocols, involving field blanks, standard additions, certified reference materials, etc.

Opportunities for harmonisation?

During this EQSPAE 2011 workshop, several facts became clear:

- There are countries with well-established and long-running WQG programmes;
- Many countries have recently established WQG programs;
- Priority substances lists for WQG development are very similar across jurisdictions;
- Manpower and financial resources for WQG development are limited in every jurisdiction;
- Terminology differs across jurisdictions, but the underlying meaning is often comparable; and
- Toxicity data evaluation and categorization criteria are similar across many jurisdictions.

Therefore, jurisdictions are duplicating each other's efforts, and more could be achieved through co-operation in WQG development. There is interest and potential for collaboration and the sharing of information, expertise, experience and best practices between jurisdictions. Ideally, jurisdictions would work together to derive international WQGs for priority substances. While this is recognized as the ultimate goal, different jurisdictions have different legal requirements, regulatory settings and environmental protection goals precluding the derivation of universally acceptable WQGs in the near future. Superimposed upon the political and technical challenges of WQG derivation are also the different regulatory speeds at which jurisdictions move towards prioritising environmental protection and accounting for changes in scientific understanding of chemical fate and behaviour. However, this should not prevent collaboration and harmonisation of frameworks, or parts thereof, for deriving WQGs.

For example, the first element in this process could be a central repository for sharing data from WQG-developing jurisdictions. Preferably, an international agreement could be developed to ensure that data are contributed and centrally managed. A less effective means would be to do this on a voluntary basis. The information collected in the data repository should probably include (but not be limited to):

- Participating jurisdictions and their relevant contact information;
- WQG derivation techniques (i.e. jurisdictional protocols);
- Findings and information related to WQG derivation methods (i.e. supporting studies to methods, summaries of development and derivation experiences, validation studies, etc.);

- Established WQGs from participating jurisdictions; and
- Standardised ecotoxicological studies suitable and used in the derivation of the respective WQG for a substance with jurisdictional evaluation report, and also those data rejected as not being suitable and reasons for the rejection (e.g. UK Technical Advisory Group on the Water Framework Directive 2008).

Participants in this workshop suggested that a mixture of people from government, academia, industry and consultants from various backgrounds related to water quality management would be required to develop international links and facilitate sharing of 'best practice' on how to develop WQGs. An International Expert Working Group could also be formed to enhance capacity building and develop an agreed protocol for deriving international WQGs. Members of the team would need to have extensive experience in WQG development and, ideally, in jurisdictional protocol development.

The first step could be an international workshop on Protocol Development to discuss issues such as (1) the choice of approaches and methods (e.g. probabilistic verses deterministic); (2) suitable toxicity studies (i.e. field, mesocosm and laboratory-based studies); (3) toxicity data evaluation; (4) choice and representativeness of ecosystem-specific and region-specific species and (5) ecosystem-specific WQG development protocols for different aquatic ecosystems (e.g. coastal and marine, estuary, freshwater lakes, rivers, man-made reservoirs, inland saline systems, etc.).

The second step could be to setup a web-based database system (for example, that might be called ATERA—Aquatic Toxicology and Ecological Risk Assessment) which would resemble the GenBank® and enable the International Expert Working Group to collect all publicly available ecotoxicological data through voluntarily uploading information by the data producers and, at the same time, to allow users to download relevant information for WQG development or ecological risk assessment. The deposited information should include, but not limited to, the toxicity test descriptions [e.g. test types (i.e. lab-based acute/chronic tests or mesocosom- or field-based tests), species name, species origin and its geographical range, test duration, toxicity endpoints, test solution system (i.e. static, semi-static-renewal or flow-through system), test conditions (i.e. temperature, hardness, salinity, pH, photoperiod, etc.), single vs. mixture chemical exposure, quantification of chemical concentrations in test solutions, references to the standard protocols adopted, etc.], all raw toxicity data in a standard format and quality assurance or quality control procedures if any. After submission of the data, the producers of the data will be formally recognised and acknowledged in the database system, for example, users can cite the reference number assigned to each dataset and authors of a particular dataset downloaded from the database. Since the data limitation represents a major bottleneck for WQG derivation especially for newly emerged chemicals, this novel web-based platform if established will facilitate rapid advancement in ecotoxicology and water quality management around the globe. To further provide more incentives to researchers, it is also possible to consider concurrently establishing a peer-reviewed electronic journal (for example that might be called *Ecotoxicological Resources*) which will primarily publish ecotoxicology data and new methods for toxicological data analysis. Such datasets will also be uploaded onto the ATERA database. A relevant example of such kind of data-resource-based journals would be *Molecular Ecology Resources*. However, it is important to acknowledge some of the clear challenges associated with such a voluntary exercise (Costa Silva and Dubé 2013) and to understand that even with clear guidance and the same starting point (in terms of data), very different WQGs can be generated by different jurisdictions (Junghans et al. 2012).

Ultimately, the International Expert Working Group should decide how far international harmonization is possible. While the ultimate goal would be the development of international Eco-region/Geo-region WQG values accepted and used by the appropriate jurisdictions in the respective Eco-/Geo-regions (i.e. internationally binding values), this might currently not be possible. Therefore, the International Expert Working Group may have to develop a protocol that allows the harmonized WQG development up to a certain point [e.g. data compilation and evaluation (= dataset development), and outlining recommended methods for WQG value setting], while the final step is left to the respective jurisdictions (i.e. selecting the derivation method and setting the actual guideline value). Or, the protocol outlines the derivation of internationally recommended WQG values, while the respective jurisdictions are free to adopt or modify the recommended values. Although some notable advancement of international collaboration in this important endeavour have been seen in Europe, the Association of Southeast Asian Nations, and Australia and New Zealand, more international collaborative efforts are urgently needed in other parts of the world (e.g. East Asia, Africa and South America) to improve the rate of progress in the derivation and application of WQGs for better management of both legacy and emerging chemicals in aquatic ecosystems worldwide.

The movement towards enhanced international harmonisation will of necessity lead to increased standardisation of methodologies. This has a number of advantages such as increased comparability of results, data having a certain minimum level of scientific rigour and increased scientific benefit for reduced effort. But conversely standardising ecotoxicology methods and the development of organisations that control, regulate and develop approved methods can have a perverse outcome. Increased standardisation has the potential to actively discourage the development of new methods (as they are not standardised and therefore cannot be used to derive WQGs), and it is likely to dramatically slow down the adoption and implementation of new advances. Standardisation is thus analogous to the movement of glaciers compared to that of running water. At the same time, a high degree of certainty is required before adopting new developments into the derivation of WQGs because failing to provide an adequate degree of protection could have major and potentially irreversible implications for the affected environments. It is therefore crucial when developing frameworks for deriving WQGs that the right balance is achieved between caution and adopting new scientifically rigorous studies and data.

Conclusions

Many regulatory jurisdictions have well-developed protocols and guidance for the derivation and implementation of WQGs for chemicals. While terminology often differs between jurisdictions, the majority of the process, types of ecotoxicity data used and the methods of WQG derivation are similar. Nevertheless, it is imperative that these protocols and guidance are regularly reviewed in order to account for developments in our understanding of the behaviour and fate of chemicals in the aquatic environment across a global scale. For example, there is a need for better and more appropriate use of available data, including the choice of summary statistic from test data to use in the derivation process, to the inclusion of higher-tier data in weight-of-evidence approaches. Some jurisdictions avoid the use of these latter types of data because of complexity of interpretation, but these data often represent a step towards the ecosystem reality that is the protection objective.

Through validation of the WQG, we can ensure that it achieves the desired level of protection in the field that is imperative especially if the WQG is enshrined in legislation. However, such a validation is currently rarely undertaken, and if it becomes mandatory, there is a need for a clear and unambiguous set of criteria by which to make the assessment of success. Methods for interpreting monitoring data for setting discharge permits and compliance assessments are well developed across many jurisdictions. Nonetheless, dealing with censored datasets and the requirements of biota monitoring to meet compliance needs for hydrophobic and bioaccumulative chemicals should be areas of cross jurisdictional development in the near future.

Our understanding of managing the challenges chemicals in the aquatic environment has developed greatly over the last 10 years. Nevertheless, it is often the case that our WQG derivation practices and guidance do not reflect these changes (e.g. incorporation of bioavailability for metals). Through closer working practices between jurisdictions and international harmonisation of WQG derivation methods or parts thereof, it is possible that knowledge can be shared and efforts and data pooled in order to minimise duplication of effort and thus permit the tackling of more recent challenges in WQG derivation and implementation, such as dealing with validation, endocrine-disrupting chemicals, mixtures and chemical metabolites. However, the challenges of harmonising approaches should not be underestimated.

Acknowledgments This work and the EQSPAE-2011 Conference are largely supported by Environment and Conservation Fund of the Government of the Hong Kong Special Administrative Region of the People's Republic of China (ECF Project 5/2011 to KMY Leung), and partially supported by the State Key Laboratory in Marine Pollution (based at City University of Hong Kong), and School of Biological Sciences of the University of Hong Kong (HKU). The authors specially thank the following workshop participants, who had joined the discussion and shared their valuable views on the scientific derivation and application of WQGs, included Put Ang, Tae-Seob Choi, Kevin K. Y. Ho, Ryan H. L. Ip, Edward T. C. Lau, Hai-Yan Li, Myung-Sook Jung, Jing-Li Mu, Sunny Sun, Greta Tam, Thiyagarajan Vengatesen, Zhen Wang, Feng-Chang Wu, Elvis G. B. Xu, Jie Xu, Andy X. L. Yi and Jian-Gang Zhu. The authors are also thankful to the staff of HKU, in particular, Helen Leung and Janet Chan for their great assistance for making the conference and workshop a great success. Apart from the first and the last authors, the order of authorships follows an alphabetical order.

References

Alberta Environmental Protection (1996) Protocol to Develop Alberta Water Quality Guidelines for Protection of Freshwater Aquatic Life. Available at http://environment.gov.ab.ca/info/library/7006.pdf. Accessed 1 Dec 2012

An YJ, Lee JK, Cho S (2008) Korean water quality standards for the protection of human health and aquatic life. Proceedings of the 2nd International Forum on Water Environment Partnership in Asia. Paper No. 10. Available at http://www.wepa-db.net/pdf/0712forum/paper10.pdf. Accessed Dec 2012

ANZECC and ARMCANZ (2000a) Australian and New Zealand guidelines for fresh and marine water quality. National Water Quality Management Strategy Paper No 4, Australian and New Zealand Environment and Conservation Council/Agriculture and Resource Management Council of Australia and New Zealand, Canberra

ANZECC and ARMCANZ (2000b) Australian Guidelines for water quality monitoring and reporting. National Water Quality Management Strategy Paper No 7, Australian and New Zealand Environment and Conservation Council / Agriculture and Resource Management Council of Australia and New Zealand, Canberra

CCME (2007) 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 (CCME), Winnipeg, Canada

Chambers PA, Culp JM, Glozier NE, Cash KJ, Wrona FJ, Noton L (2006) Northern Rivers Ecosystem Initiative: nutrients and dissolved oxygen—issues and impacts. Env Mon Ass 113:117–141

Chapman PM, McDonald B, Kickham PE, McKinnon S (2006) Global geographic differences in marine metals toxicity. Mar Pollut Bull 52:1081–1084

Costa Silva G, Dubé MG (2013) Water quality assessment at a global scale: a comparison between world regions. Water Internat 38:78–94

Crane M, Kwok KWH, Wells C, Whitehouse P, Lui GCS (2007) Use of field data to support European Water Framework Directive quality standards for dissolved metals. Environ Sci Technol 41:5014–5021

Crane M, Matthiessen P, Maycock DS, Merrington G, Whitehouse P (2010) Derivation and use of environmental quality and human health standards for chemical substances in water and soil. CRC Press, Boca Raton, p 140

Crommentuijn T, Polder MD, Posthumus R, Van de Plassche E (1997) Maximum permissible concentrations and negligible concentrations for metals, taking backgrounds concentrations into account. Netherlands Institute of Public Health and the Environment, RIVM Report No. 601501001, Bilthoven, the Netherlands

EC (European Commission) (2011) Guidance Document No. 27. Technical Guidance for Deriving Environmental Quality Standards. Technical Report-2011–055. European Communities, Brussels. Available at http://circa.europa.eu/Public/irc/env/wfd/library?l=/framework_directive/guidance_documents/tgd-eqs_cis-wfd/_EN_1.0_&a=d. Accessed 1 Dec 2012

EC (European Commission) (2012) Proposal for a Directive of the European Parliament and of the council amending Directives 2000/60/EC and 2008/105/EC as regards priority substances in the field of water policy. COM(2011) 876 final. Brussels, Belgium

Environment Agency (2008) Determination of metal background reference concentrations: feasibility study. Environment Agency Science Report SC050063/SR, Bristol, UK

Environment Agency (2009) Using biotic ligand models to help implement environmental quality standards for metals under the Water Framework Directive. Science Report SC080021/SR7b, Environment Agency, Bristol, UK

FMWQ Working Group 4 (2010) Draft FMWQ Working Group 4 Workshop Minutes. Fresh and Marine Water Quality Working Group 4—toxicants and sediments. 14–16 April 2010. Lucas Heights, NSW, Australia

Forbes TL, Forbes VE (1993) A critique of the use of distribution-based extrapolation models in ecotoxicology. Function Ecol 7: 249–254

Fox DR (2010) A Bayesian approach for determining the no effect concentration and hazardous concentration in ecotoxicology. Ecotoxicol Environ Safe 73:123–131

Gronewold AD, Borsuk ME (2010) Improving water quality assessments through a hierarchical Bayesian analysis of variability. Environ Sci Technol 44:7858–7864

Hamon RE, McLaughlin MJ, Gilkes RJ, Rate AW, Zarcinas B, Robertson A, Cozens G, Radford N, Bettenay L (2004) Geochemical indices allow estimation of heavy metal background concentrations in soils. Global Biogeochem Cycle 18:GB1014:1–6

Helsel DR (2005) More than obvious: better methods for interpreting nondetect data. Environ Sci Technol 39:419A–423A

Helsel DR, Cohn T (1988) Estimation of descriptive statistics for multiply censored water quality data. Water Resour Res 24:1997–2004

Hobbs DA, Warne MStJ, Markich SJ (2004) Utility of northern hemisphere metal toxicity data in Australasia. SETAC Globe 5(2): 38–39

Hose GC (2005) Assessing the need for groundwater quality guidelines using the species sensitivity distribution approach. Human Ecol Risk Assess 11:951–966

International Organisation for Standardisation (2005a) ISO 19258: 2005 Soil Quality: Guidance on Determination of Background Values. ICS code 13.080.99

International Organisation for Standardisation (2005b) Water quality—sampling—Part 20: Guidance on the use of sample data for decision making—compliance with limit values and classification. BSI British Standards Publications, London, p 34

Jager T (2012) Bad habits die hard: the NOEC's persistence reflects poorly on ecotoxicology. Environ Toxicol Chem 31:228–229

Jin XW, Lei BL, Xu YP, Zha JM, Wang ZJ (2009) Methodologies for deriving water quality criteria to protect aquatic life (ALC) and proposal for development of ALC in China: a review. Asian J Ecotoxicol 4:609–616

Junghans MJ, von Arb S, Whitehouse P, Johnson I (2012) Variability in environmental quality standards—how much is there and what are the causes? Poster presented at SETAC World Congress 2012, May. Berlin.

Klein JP, Moeschberger ML (2003a) Survival analysis: techniques for censored and truncated data. Springer, New York, p 536

Klein JP, Moeschberger ML (2003b) Survival analysis: techniques for censored and truncated data. Springer, New York

Kwok KWH, Leung KMY, Chu VKH, Lam PKS, Morritt D, Maltby L, Brock TCM, Van den Brink PJ, Warne MStJ, Crane M (2007) Comparison of tropical and temperate freshwater species sensitivities to chemicals: implications for deriving safe extrapolation factors. Integr Environ Assess Manag 3(1):49–67

Landis WG, Chapman PM (2011) Well past time to stop using NOELs and LOELs. Integr Environ Assess Manag 7:vi–viii

Leung KMY, Bjørgesæter A, Gray JS, Li WK, Lui GCS, Wang Y, Lam PKS (2005) Deriving sediment quality guidelines using field-based species sensitivity distributions. Environ Sci Technol 39: 5148–5156

Liney KE, Hagger JA, Tyler CR, Depledge MH, Galloway TS, Jobling S (2006) Health effects in fish of long-term exposure to effluents from wastewater treatment works. Environ Health Perspect 114: 81–89

Linton TK, Pacheco MAW, McIntyre DO, Clement WH, Goodrich-Mahoney J (2007) Development of bioassessment based benchmarks for iron. Environ Toxicol Chem 26:1291–1298

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

Meeker WO, Escobar LA (1998) Statistical methods for reliability of data. Wiley, New York

Ministry for the Environment (2003) Contaminated land management guidelines no. 2. Ministry for the Environment, Wellington

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

NEPC (National Environment Protection Council) (2011) National Environment Protection (Assessment of Site Contamination) Measure. Schedule B5b—guideline on methodology to derive ecological investigation levels in contaminated soils. NEPC, Adelaide, Australia. pp. 87. Available at: http://www.ephc.gov.au/sites/default/files/Schedule_B5b__Guideline_on_methodology_to_derive_EILs__SEP10.pdf. Accessed 20 Dec 2012

Newman MC (1995) Quantitative methods in aquatic ecotoxicology. CRC/Lewis, Boca Raton, pp 24–26

Newman MC (2008) What exactly are you inferring? A closer look at hypothesis testing. Environ Toxicol Chem 27:1013–1019

Peters A, Simpson P (2012) Accounting for both local aquatic community composition and bioavailability in setting local quality standards for metals. Platform Presentation, 6th SETAC World Congress/SETAC Europe 22nd Annual Meeting, Berlin, 20–24 May

Peters A, Crane M, Simpson P, Merrington G (2010) Assessment of the effects of nickel on benthic invertebrates in the field. Final Report Prepared for the Nickel Producers Environmental Research Association, Durham, NC, USA

Peters A, Crane M, Adams W (2011) Effects of iron on benthic macroinvertebrate communities in the field. Bull Environ Contam Toxicol 86:591–595

Posthuma L, Suter GW, Traas TP (eds) (2001) Species sensitivity distributions for ecotoxicology. Lewis Publishers, Boca Raton

SCHER (Scientific Committee on Health and Environmental Risks) (2010) Opinion on the Chemicals and the Water Framework Directive: technical guidance for deriving environmental quality standards. Available at http://ec.europa.eu/health/scientific_committees/environmental_risks/docs/scher_o_127.pdf. Accessed 1 Dec 2012

Singh A, Nocerino J (2002) Robust estimation of mean and variance using environmental datasets with below detection limit observations. Chemometrics Intellig Lab Syst 60:69–86

Stephan CE, Rogers JW (1985) Advantages of using regression analysis to calculate results of chronic toxicity tests. In: Bahner RC, Hansen DJ (eds) Aquatic toxicology and hazard assessment. STP 891. American Society for Testing and Materials, Philadelphia, pp 328–338

Stephan CE, Mount DI, Hanson DJ, Gentile JH, Chapman GA, Brungs WA (1985) Guidelines for deriving numeric National Water Quality Criteria for the protection of aquatic organisms and their uses. U.S. Environmental Protection Agency Report PB85-227049, Washington, DC, USA

UK Technical Advisory Group on the Water Framework Directive (2008) Proposals for Environmental Quality Standards for Annex VIII Substances. Available at http://www.wfduk.org/sites/default/files/Media/Environmental%20standards/Specific%20pollutants%20proposals_Final_010608.pdf. Accessed 1 Dec 2012

US EPA (1994) Guidance for the data quality objectives process. EPA QA/G-4. EPA/600/R-96/055

US Environmental Protection Agency (1985) Guidelines for deriving numerical national water quality criteria for the protection of aquatic organisms and their uses. Available at http://water.epa.gov/scitech/swguidance/standards/criteria/aqlife/index.cfm#guide. Accessed 1 Dec 2012

Van Sprang P, Delbeke K, Regoli L, Waeterschoot H, Van Assche F, Adams W, Haesaerts D, Mattelet C, Bush A, Chung L, Verougstraete V (2008) Assessment of metal bioavailability and natural background levels—WFD monitoring from the perspective of metals industry. In: Qusvauviller P, Borchers U, Thompson C, Simonart T (eds) The Water Framework Directive: ecological and chemical status monitoring. Wiley, New York, pp 299–312

Warne MStJ (1998) Critical review of methods to derive water quality guidelines for toxicants and a proposal for a new framework. Supervising Scientist Report 135, Supervising Scientist, Canberra, ACT, Australia. ISBN 0 642 24338 7. 82 p

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

Wepener V, Chapman PM (2012) South African ecotoxicology—present status and future prognosis. African J Aquat Sci. doi:10.2989/16085914.2012.717051

Wheeler JR, Grist EPM, Leung KMY, Morritt D, Crane M (2002) Species sensitivity distribution: data and model choice. Mar Pollut Bull 45:192–202

Wu FC, Meng W, Zhao XL, Li HX, Zhang RQ, Cao YJ, Liao HQ (2010) China embarking on development of its own national water quality criteria system. Environ Sci Technol 44:7992–7993

Zhao FJ, McGrath SP, Merrington G (2007) Estimates of ambient background concentrations of trace metals in soils for risk assessment. Environ Pollut 148(1):221–229