



Australian Government



AUSTRALIAN INSTITUTE
OF MARINE SCIENCE

KEY ISSUES IN THE DERIVATION OF WATER QUALITY GUIDELINE VALUES: A WORKSHOP REPORT

Rebecca Fisher, Rick van Dam, Graeme Batley, David Fox, Andrew Harford, Chris Humphrey, Cath King, Patricia Menendez, Andrew Negri, Abigail Proctor, Quanxi Shao, Jenny Stauber, Joost van Dam and Michael Warne



AIMS: Australia's tropical marine research agency.

www.aims.gov.au

Australian Institute of Marine Science

PMB No 3
Townsville MC, Qld 4810

PO Box 41775
Casuarina, NT 0811

Indian Ocean Marine Research Centre
University of Western Australia, M096
Crawley, WA 6009

This report should be cited as:


Fisher R, van Dam R, Batley G, Fox D, Harford A, Humphrey C, King C, Menendez P, Negri A, Proctor A, Shao Q, Stauber J, van Dam J and Warne M. (2019). Key issues in the derivation of water quality guideline values: a workshop report. *Australian Institute of Marine Science Report, Crawley, WA, Australia. (51 pp).*

© Copyright: Australian Institute of Marine Science (AIMS) [2019]

All rights are reserved and no part of this document may be reproduced, stored or copied in any form or by any means whatsoever except with the prior written permission of AIMS

DISCLAIMER

While reasonable efforts have been made to ensure that the contents of this document are factually correct, AIMS does not make any representation or give any warranty regarding the accuracy, completeness, currency or suitability for any particular purpose of the information or statements contained in this document. To the extent permitted by law AIMS shall not be liable for any loss, damage, cost or expense that may be occasioned directly or indirectly through the use of or reliance on the contents of this document.

Project Leader shall ensure that documents have been fully checked and approved prior to submittal to client				
Revision History:		Name	Date	Comments
1	Prepared by:	Rebecca Fisher 	27/08/2019	
	Reviewed by:	Ross Jones	27/8/2019	
	Approved by:	Richard Brinkman	27/08/2019	
2	Prepared by:	<i>Insert Lead Author's name</i>	<i>Insert date</i>	
	Reviewed by:	<i>Insert Reviewer name</i>	<i>Insert date</i>	
	Approved by:	<i>Insert Program Director's name</i>	<i>Insert date</i>	

Cover photo:

SeaSim, Australian Institute of Marine Science, Cape Cleveland Australia. Image: C. Miller

CONTENTS

EXECUTIVE SUMMARY.....	iv
1 INTRODUCTION.....	1
1.1 Workshop objectives	1
1.2 Key issues	2
2 SSD MODEL DISTRIBUTIONS AND MODEL FITS	4
2.1 Species sensitivity distribution fitting software	4
2.1.1 Burrlioz.....	4
2.1.2 The ‘ssdtools Shiny app’	5
2.2 Model averaging.....	6
2.3 Bimodality	7
2.4 Focussing on the tails	7
2.5 Goodness of fit	8
2.6 Comparing SSDs and GVs	8
2.7 Recommendations for testing SSD fitting methods and the ssdtools Shiny app.....	9
3 ACCOUNTING FOR UNCERTAINTY IN GUIDELINE VALUE DERIVATION	11
3.1 Capturing uncertainty in underlying toxicity data.....	11
3.2 Incorporating all toxicity estimates for a single endpoint.....	12
3.3 Dealing with other sources of uncertainty.....	13
3.4 How should uncertainty and variability be used?	14
3.5 Recommendations for capturing and reporting uncertainty in GV derivation	15
4 INPUT TOXICITY DATA FOR SSDs.....	16
4.1 Toxicity estimates	16
4.1.1 Estimating the no effect concentration (NEC).....	16
4.1.2 General considerations	18
4.2 Acute versus chronic endpoints.....	19
4.3 Other temporal considerations	20
4.4 Reliability of ecotoxicity data and use of poor-quality ecotoxicity data.....	21
4.5 Species selection.....	23
4.5.1 Sensitive species or ‘keystone’ species?	24
4.6 Recommendations regarding input toxicity data.....	25

5	WEIGHTING OF INDIVIDUAL DATA POINTS / DATASETS	26
5.1	Weighting based on data quality	26
5.2	Weighting based on taxonomic representation	26
5.3	Recommendations	27
6	SMALL DATASETS.....	28
6.1	Recommendations	29
7	ALTERNATIVES TO SSDs	30
7.1	Probabilistic thresholds.....	30
7.1.1	Can a 'true' PC99, or even PC95 actually be estimated?.....	30
7.1.2	How can PCx values be directly linked to environmental and economic cost?	31
7.1.3	Probabilistic threshold derivation	31
7.2	Multivariate methods.....	32
7.2.1	Similarity based community analysis	32
7.2.2	TITAN.....	34
7.3	Recommendations	34
8	SUMMARY.....	35
9	REFERENCES	38
	APPENDIX A: WORKSHOP ATTENDEES.....	42
	APPENDIX B: KEY ISSUES	43
	SSD model distributions and model fits.....	43
	Uncertainty in GV derivation.....	43
	Input toxicity data for SSDs.....	44
	Derivation and selection of toxicity estimates	44
	Temporal considerations	44
	Reliability of data and use of poor quality ecotoxicity data.....	44
	Weighting of individual data points/datasets	44
	Taxonomic/regional ecological relevancy of the datasets.....	44
	Small datasets.....	45
	Alternatives to SSDs.....	45
	How to we integrate/incorporate other information (lab/field/expert opinion)	45
	Exposure mechanisms/characteristics	45
	Implementation/application of guidelines.....	46
	Statistical practice in ecotoxicology.....	46

LIST OF TABLES

Table 1. Topics and sub-topics raised in the “key issues” survey by workshop participants and their number of received votes.....	2
Table 2. Key issues identified by respondents relating to the Burrlioz software.....	4
Table 3. Key issues identified by respondents relating to the use of poor quality data and/or relevance of endpoints.....	23
Table 4. Key issues identified by respondents relating to species selection and taxonomic representativeness.....	24

EXECUTIVE SUMMARY

A technical workshop was held from March 27–29, 2019 at the Australian Institute of Marine Science (AIMS) in Townsville, Queensland, to discuss key issues associated with the derivation of water quality guideline values (GVs). The workshop arose following discussions in 2018 amongst some of the workshop attendees on potential improvements to current GV derivation methods, and an associated presentation at the SETAC Europe conference in Rome, in May 2018. The workshop was funded by a “Community of Practice” (CoP) grant from AIMS. Fourteen participants attended, comprising ecotoxicologists, environmental chemists, ecologists and biostatisticians (Appendix A). A pre-workshop survey was distributed to the workshop invitees as well as to numerous external parties, seeking views on the most pressing issues/questions facing GV derivation. The workshop structure was determined by the attendee group at the outset of the workshop after considering and prioritising the issues raised in the survey.

This report summarises the discussions of the workshop and outlines the necessary steps to improve the GV derivation methods to maximise their benefit to stakeholders in the scientific, regulatory and industrial sectors. A particular outcome was the desirability of pursuing already initiated collaborations with Canadian scientists with a view to enhancing and adopting their model averaging approach to GV derivations using ssdtools Shiny app recently developed for the British Columbia Ministry of the Environment. For each topic discussed, a range of recommendations was included regarding the best approach for exploring, progressing and/or improving the use of statistical methods in ecotoxicology and water quality guideline derivation.

1 INTRODUCTION

A technical workshop was held from March 27–29, 2019 at the Australian Institute of Marine Science (AIMS) in Townsville, Queensland, to discuss key issues associated with the derivation of water quality guideline values (GVs). The workshop arose following discussions in 2018 amongst some of the workshop attendees on potential improvements to current GV derivation methods, and an associated presentation at the SETAC Europe conference in Rome, in May 2018. The workshop was funded by a “Community of Practice” (CoP) grant from AIMS. Fourteen participants attended, comprising ecotoxicologists, environmental chemists, ecologists and biostatisticians (Appendix A). A pre-workshop survey was distributed to the workshop invitees as well as to numerous external parties, seeking views on the most pressing issues/questions facing GV derivation. The workshop structure was determined by the attendee group at the outset of the workshop after considering and prioritising the issues raised in the survey.

1.1 Workshop objectives

The initial intent was to:

(i) Address issues and solutions with the currently practiced species sensitivity distribution (SSD) methods

- Review the currently practiced statistical methods in ecotoxicology with respect to GV derivation, including:
- Identification of the key statistical issues and knowledge gaps associated with the current methods
- Identification of existing methods that address those issues and/or fill those knowledge gaps
- An evaluation of whether the existing methods are fit-for purpose and/or best practice in the context of practical application
- For existing fit-for purpose methods currently not adopted, assess the reason for lack of adoption and develop an action plan for implementation/adoption
- Identification of statistical issues and knowledge gaps NOT currently addressed using fit-for-purpose methods
- Discuss/explore and agree on the most likely fit-for-purpose approaches to address outstanding issues and gaps.

(ii) Consider alternative derivation methods

- Invite attendees to propose and present ideas on alternative derivation methods (especially for low sample sizes) for group discussion with the aim of reaching a consensus on a set of promising approaches that can then be put to further test.

1.2 Key issues

Prior to the workshop each attendee was asked to provide a list of 5 to 10 key questions/problems/issues associated with current methods used for the derivation of water quality guideline values (also referred to as standards or criteria). All responses from the pre-workshop survey were collated and grouped into 14 broad, inter-related sub-topics (Table 1, see the complete list of responses in Appendix B). At the start of the workshop each participant was asked to cast five votes for those sub-topics they most wished to discuss during the workshop. Three sub-topics stood out as being ranked of high importance by more than 50% of the participants. These were topics around small datasets, accounting for uncertainty in SSDs, and exploring alternatives to SSDs (Table 1). Beyond these top three topics, several others were ranked quite highly, receiving largely equal weight (Table 1). As “use of SSD confidence intervals” was a topic highly related to issues around “uncertainty in GV derivation” these were grouped for discussion. Three other topics were also selected for discussion, including issues around “SSD model distributions and model fits”, a range of sub-topics that can be broadly categorized as relating to “input toxicity data”, as well as issues around “weighting of individual data points/datasets”. Here we provide an account of discussions, and workshop participants’ views on these broad topics, flowing from issues specific to SSD model distributions and model fitting, capturing and using uncertainty in derivation, input toxicity data, through to consideration of small datasets and alternatives to SSD approaches.

Table 1: Topics and sub-topics raised in the “key issues” survey by workshop participants and their number of received votes

Topic	Votes
SSD model distributions and model fits	5
Uncertainty in GV Derivation	
Capturing uncertainty	8
Use of SSD confidence intervals	5
Input toxicity data	
Derivation and selection of toxicity estimates	5
Acute versus chronic endpoints and other temporal considerations	4
Reliability of data and use of poor-quality ecotoxicity data	2
Taxonomic issues	4
Weighting of individual data points/datasets	5
Small datasets	9
Alternatives to SSDs	8
Other topics	
How to integrate/incorporate other information (lab/field/expert opinion)	5
Exposure mechanisms/characteristics	2
Implementation/application of guidelines	0
Statistical practice in ecotoxicology	0

Time constraints precluded the inclusion of discussion around a range of other sub-topics, although some of these were certainly considered of importance (e.g. “how to integrate/incorporate other information”, Table 1).

2 SSD MODEL DISTRIBUTIONS AND MODEL FITS

The Australian and New Zealand Guidelines for Fresh and Marine Water Quality (ANZG, 2018), preferentially uses a species sensitivity distribution (SSD) approach to derive guideline values (GVs) that are protective or hazardous to a given percentage of species (PCx), although an assessment factor approach can also be used if there are insufficient data for the SSD approach. In the SSD approach, the cumulative probability distribution of the observed chronic toxicity endpoint data for a range of aquatic species is modelled (using a statistical distribution) and used to interpolate (or extrapolate) the concentration that protects a specified percentage of species, often termed a protective concentration (PC; e.g. PC95 = the concentration predicted to protect 95% of species) (ANZG, 2018; Warne et al., 2018). Depending on the level of protection required for an ecosystem, a specific PC is selected as a GV.

2.1 Species sensitivity distribution fitting software

2.1.1 Burrlioz

The currently accepted method for fitting SSD models for deriving GV in Australia and New Zealand is through the use of the Burrlioz 2.0 software (Barry and Henderson, 2014). Many of the survey respondents raised issues around the use of this software from their own experience with fitting SSD curves. Some specific ‘bugs’ and limitations identified by respondents for Burrlioz 2.0 software are listed in Table 2. Some of the issues raised reflect a lack of awareness (i.e. there is in fact a user guide that contains the statistical details, including the code), or a lack of understanding (AIC provides only a relative measure of fit) of some users. Regardless, it was clear that there is a high level of frustration among the user community around the use of the Burrlioz 2.0 software, and that an update or recommendations of alternative software warrants consideration. While the underpinnings of the BurrliOz software are ‘open source’ as it is based on the open-source, freely available software R, and all the code used to fit the included SSD relationships is provided in the various documentation, the actual software itself is not, and can only be modified by the developers (CSIRO).

Table 2: Key issues identified by respondents relating to the Burrlioz software

Differences in model fits (but no differences in the PCx values) when using the same data in Burrlioz 2.0 compared to the earlier version (Burrlioz 1.0) due to differences in how the y axis plotting positions are calculated, which makes it impossible to compare the fit of the SSDs between software versions.
--

On occasions, the statistical distribution could not be fitted to the data unless the data values were multiplied by a factor such as 10 and then the resulting PCx divided by 10.
--

Getting poor fits due to the limitations of the models that Burrlioz 2.0 can apply, as compared to the SSD software recently developed by British Columbia, Canada, that fits multiple distributions and derives the estimates based on a weighted average of the fits of the distributions (see more discussion below on model averaging and ssdtools).
Software cutting off data points in the Burrlioz report print out when there are greater than approximately 30 data points, or cutting off plot labels if species name is too long.
The Burrlioz 2.0 interface is sometimes difficult to work with, e.g. you can't scroll down on the front page, there can be issues with loading a new set of data after completing one SSD, so have to close and open software and lose all graphic settings, and labelling of x axis with symbols is difficult unless you know R language.
No help/guidance option in the software providing information on how the software fits the models, e.g. does it force the model through zero which will affect lower PC ₉₉ values or is it asymptotic to infinity? How are the 95% confidence limits on the PC ₉₅ etc calculated? How are the confidence bands on the SSD calculated – are they CIs or prediction intervals? It is currently a “black box” interface where numbers go in and model comes out. One option for providing a fit on the model to the data would be to calculate the Akaike Information Criterion (AIC).
Lack of technical support in maintaining and updating Burrlioz 2.0.
The current software for deriving SSDs in Australia does not always fit the data well – can its flexibility be improved?
The ability to fit multiple different models and choose the best fit would be an improvement on the current situation.
Restricting the choice of distributions in Burrlioz might not be offering the best predictions, especially for small datasets, so we should start including other model types?
Burrlioz fits have been identified as particularly poor in fitting the head and/or tail of the distribution for datasets with a large concentration range (>4 orders of magnitude), and when the concentration data are ng/L.

2.1.2 The ‘ssdtools Shiny app’

A potential alternative to Burrlioz 2.0 is the ssdtools Shiny app (Dalgarno, 2018, 2019) built by Poisson Consulting for the Ministry of the Environment, British Columbia, and available at: <https://poissonconsulting.shinyapps.io/ssdtools/>. Shiny is a web-based tool (Chang et al., 2019) for building user friendly web interfaces (applications or ‘apps’) to provide greater accessibility to the numerous ‘R’ packages available for various statistical analyses. The Shiny platform shows considerable promise in the context of developing a range of user-friendly interfaces to allow wider adoption of ‘best practice’ statistical methods across the full water quality derivation process. The Poisson Consulting Shiny app is based on the ‘R’ package ‘ssdtools’ (Thorley and Schwarz, 2018). The ssdtools Shiny app was discussed at length at the workshop, and the collective opinion was that it should be investigated as a potential alternative to Burrlioz 2.0. The ssdtools Shiny app interface is intuitive and easy to use, although thorough testing is advised with respect to some of the identified ‘bugs’ of Burrlioz. The open source nature of the application, and the fact that the tool not only produces the R code used to generate a specified model fit, but that all the underlying code are freely available on the Province of British Columbia github.com site (<https://github.com/bcgov/ssdtools>) ensures that the methods used are completely transparent and easily replicable.

2.2 Model averaging

The default approach in the ssdtools Shiny app is to fit six candidate distributions and provide model-averaged PC_x values to the user. Model averaging is a procedure that stems from an information theoretic framework, which acknowledges that there is uncertainty in which model is 'true', and instead bases inference on all of the models in a candidate set, using model weights calculated through Akaike Information Criteria (AIC, Burnham and Anderson, 2002). The ssdtools Shiny app approach uses a weighted geometric mean, based on a modification of the AIC to account for small sample sizes (AICc, Burnham and Anderson, 2004). Use of AICc over AIC is appropriate, given the inherently small samples available for most SSD-based GV derivations.

There are different ways that model averaging can be achieved in practice, using AICc model weights. Each model can be fitted and used to estimate a given PC_x value, with each PC_x value being weighted according to the relative fit of the model from which it was estimated. The model-averaged PC_x value is a weighted average of individual PC_x values. Alternatively, a *mixture* of probability distributions can be fitted to the data from which the desired PC_x values can be estimated. Both approaches (model averaging and mixture modelling) are similar in spirit, although computationally different. Furthermore, whereas any number of distributions can (conceptually at least) be used in the averaging process, the same is not true of mixture-modelling. Fox (*pers com.*) recommends that the number of data points used to fit the mixture-model be at least 3 times the total parameter count of the distribution. Applying this 'rule' suggests that a mixture-model (involving two 2-parameter distributions, with a mixing parameter, i.e. 5 parameters) should only be contemplated for sample sizes of 15. This raises another issue which requires closer examination and that concerns co-dependency among the parameter estimates when using model-averaging for small data sets. For example, a model-averaged PC_x based on five, 2-parameter distributions fitted to $n=8$ data points would appear to be an abuse of the process. Further testing is required to assess the extent to which these different approaches may differ in their estimates of the PC_x values (and associated uncertainty), and how these methods may perform in the case of small sample sizes.

Model averaging is controversial in some settings, as some argue that the best model should be selected on a theoretical basis and not driven entirely by statistics. The main advantage of model-averaging and mixture-modelling is the ability to allow a number of theoretically plausible models to be fitted, with those showing the strongest fit to the observed data weighting higher in the resulting model averaged PC_x values. Further, mixture-modelling is commensurate with the notion that different toxicity metrics are described by fundamentally different probability models. Both approaches address two key issues identified with the currently used methods. Firstly, they avoid the identification of a single 'best' distribution. Secondly, both approaches provide greater flexibility in the fitted model, which should allow a wider range of responses to be modelled. There was general consensus at the workshop that both model-averaging and mixture-modelling may prove to be a robust general solution to some of the problems associated with SSD modelling, where there is no sound theoretical basis for selection of one functional model form over another. However, there are some concerns that warrant investigation before such an approach should be broadly adopted.

2.3 Bimodality

The current statistical approach for the derivation of SSDs does not accommodate bimodal or multimodal datasets. The sigmoidal distributions that are applied using, for example, the BurrIIOs and ssdtools Shiny app software do not accommodate non-sigmoidal SSDs. As a consequence, SSDs for many pesticides (e.g. PSII herbicides, neonicotinoid insecticides) need to be based on a cut down toxicity dataset that includes data for only the most sensitive trophic levels (i.e. phototrophic organisms in the case of PSII herbicides), and therefore, the SSDs represent only a subset of the community. Currently, the issue of bi-modality is addressed through the use of taxa-specific SSDs with a focus on the most sensitive group, as this represents a precautionary approach that might be particularly appropriate for highly toxic compounds. However, there are instances where the ability to apply mixture distributions that can accommodate bi-/multimodal distributions may be of value.

David Fox has undertaken some preliminary investigations into whether the model-averaging approach may resolve some of the multi-modality issues, by fitting a mixture of distributions. The outcome of this testing suggests that model predictions in the case of extreme bi-modality are not resolved through the use of AICc weighted model averaging, and that this can only be effectively modelled using a formal mixture model framework, where estimation of the mixing proportion is explicitly estimated during the model fitting.

It may be possible to incorporate models that account for bimodality into the model-averaging approach, providing a suitable set of potential models was formulated, and fitting algorithms developed and made available. One candidate model is the mixture model used by David Fox in testing of the performance of model-averaged predictions in the case of extreme bimodality, or that of Zajdlik et al. (2009). Bi- and multi-modal models have substantially more parameters than uni-distributional models, which raises the issue: how do we balance models with greater numbers of parameters when datasets used to construct SSDs are often small? Rigorous testing would be required to assess the performance of AICc in generating model weights that appropriately reflect the level of 'real' support any given dataset is expected to have for a more complex model, given the sample size and distribution of the underlying SSD endpoint data.

2.4 Focussing on the tails

The fact that SSD curves are most commonly used to estimate 'conservative' GVs representing very high levels of species protection (e.g. PC95 and PC99) raises the following questions: (i) Should we be focusing more on getting a reliable fit at the bottom end of the SSD? (ii) Are there any other ways to improve the reliability at the bottom end of the distribution, given the strong focus on high levels of protection? On the other hand, SSDs are meant to be a representation of the sensitivity of all species of interest in the environment, not just the sensitive species and, thus, the distribution needs to be representative of the overall dataset. Statistically, you must assess the fit overall because that is what the probability model is attempting to describe and focussing or restricting the computations and assessment to just a very small portion of the distribution will not guarantee that the resulting distribution qualifies a true pdf, in which case all subsequent estimation and inference is discarded.

One possibility for focusing the fit of the SSD distributions on the tail is to use a ‘segmented’ or ‘truncated’ distribution, such as the half-t or half-normal, as has been implemented by Proctor (2018). With the use of these segmented distributions it is possible to fit a distribution using a given percentage of the most sensitive data (e.g. all data below the median). In doing this, you are focusing the fit of the lower half distribution to the lower half of the data and assuming that the upper half of the distribution is fitted by some other, but unknown distribution (although the upper data still determine where the median falls). The logic for this practice is that the fit of the distribution to the most sensitive species in the data should not be influenced by any behaviour from data of the least sensitive species. This practice may overcome issues with bimodality and the need to use a taxa-specific SSD, and instead allows the SSD to have more focus on the fit of the lower tail, where the most sensitive species are. A criticism of this approach is that you are only using half, not all of the data, not normally done in ecotoxicology. In addition, the approach theoretically requires a larger sample size. Initial work with segmented distributions by (Proctor et al, PhD thesis) has shown that with the use of simulated data, segmented distributions are able to more accurately estimate PC90, 95, and 99 values compared to full distributions. More importantly, this improved fit is evident even when data is simulated to contain bimodality, or skew in the upper and lower tails of the fitted data, particularly true when the students t-distribution is used to fit the SSD.

2.5 Goodness of fit

In Australia and New Zealand, the assessment of the reliability of a GV derived using the SSD method depends on a number of criteria (e.g. sample size, data type), including the adequacy of the fit in the SSD, which is judged to be either good or poor based on best professional judgement (ideally from three experts) (Warne et al., 2018). Just as a weight of evidence approach has been introduced for assessing the modality of the data (Warne et al., 2018), a similar approach could be applied to the fit of the SSD model as well, where one line of evidence could be a statistical assessment of the fit of the SSD model. While the greater flexibility inherent in the model-averaging and mixture-modelling approaches discussed above improves the ability to incorporate aspects from a number of probability models, this does not mean that any of the candidate models considered are in fact a good fit of the data.

There is a wide range of statistical methods for assessing statistical fit of an SSD model, six of which are provided in a ‘goodness-of-fit table’ in the ssdtools Shiny app (ad, ks, cvm, aic, aicc, bic). Ideally, a goodness-of-fit measure would have some criteria by which to decide that “the predictions are too poor to be using SSD models, and other approaches should be adopted” or that the use of SSD models is appropriate. Use of AICc, AICc and BIC for this purpose is problematic, because such goodness-of-fit measures are purely relative, i.e. the absolute values for a single model have no inherent value in terms of assessing goodness of fit. These metrics are only useful in the context of comparing models.

2.6 Comparing SSDs and GVs

Use of different packages and approaches around the world to construct SSDs and derive GVs makes comparisons difficult, particularly for comparisons involving different biomes, such as temperate

versus tropical data. While global consistency in the fitting methodology of SSDs would potentially aid with SSD comparisons, such a consensus is not realistic in the short term. Any formal comparisons between SSDs fitted using data from different ‘groups’ of interest should ideally be done directly using formal statistical approaches. While not currently widely adopted in ecotoxicological statistics, there is a range of formal statistical tests for comparing the fit of two non-linear models, such as the Kolmogorov-Smirnov test that can be readily applied to comparing two SSD models. Such comparisons could also be embedded within the information theoretical framework, where AICc is used to compare a ‘model’ that allows separate parameters to be fitted across different groups of data (i.e. separate curves fitted to temperate or tropical taxa) to a ‘model’ with only a single SSD curve fitted jointly. Such an approach could provide a probability that the SSD curves from the two systems can indeed be considered different. The development of a friendly interface to carry out formal statistical comparisons of multiple SSD curves may be of considerable value to the ecotoxicological community and could potentially be achieved through Shiny app in a similar manner to the ssdtools Shiny app. Such a tool could also potentially be used to test for bimodality in the SSD data.

2.7 Recommendations for testing SSD fitting methods and the ssdtools Shiny app

Discussion at the workshop yielded the following recommendations with respect to adopting the ssdtools Shiny app approach:

- There is a need to further investigate the methods used for AICc weighting and ensure that the most appropriate method for estimating weighted PC_x values is used. For example, should all the PC_x values be estimated and a weighted average calculated (as currently implemented in the ssdtools Shiny app), or should a mixture distribution be used to estimate PC_x directly (see above).
- The models to be included in the candidate ‘model set’ need to be carefully considered. Models included must be ‘plausible’, meaning that they have the potential to provide realistic PC_x estimates and are commensurate with the nature and type of data. The Pareto model currently available in the ssdtools Shiny app would appear to be an example of a distribution that should be excluded as: (i) it is prone to stability issues in the estimation process, and (ii) it can yield non-sensical results (e.g. negative PC_x values). Furthermore, it is noted that the Pareto model is not currently included in the default set of candidate models in the ssdtools Shiny app. The models currently implemented in BurrIoz 2.0 should definitely be included, namely the log-logistic (currently one of the defaults in the ssdtools Shiny app) and Burr Type III. It was recognised that developing a set of criteria for model set selection would be beneficial.
- It would be valuable to explore how the model averaging approach is affected by sample size, for a range of example data sets. AICc appropriately heavily penalises more complex models with very small samples. The current Australian and New Zealand GV derivation methods stipulate that a log-logistic distribution be used when there are toxicity data for <8 species, and a Burr Type III distribution be used when there are toxicity data for ≥8 species (Warne et al., 2018). It is worth investigating how the model averaging approach may align with this recommendation, and if small sample sizes do in fact favour simpler models (i.e.

the log-logistic over the Burr Type III) and if so, at what sample size. Conversely, there may be some utility to including slightly more complex models (e.g. see Section 2.4 on Bimodality above) and testing that the model weights are robust with respect to ‘overfitting’ small samples. The heavy penalty against model complexity suggests that the model-averaging approach may be quite robust to the inclusion of more complex models, but this warrants thorough testing.

- Some concern was expressed that having multiple very similar models in the model set may lead to an ‘overweighting’ of that functional form in the weighted averaging. Preliminary testing suggests that where similar models are included (for example the gamma and the Gompertz) they effectively ‘share’ the model weight, suggesting that ‘overweighting’ is unlikely to be an issue, although this warrants further testing.
- For the approach to be used for deriving GVs in Australia and New Zealand, clear guidance around the ‘best practice’ approach would need to be developed, and the current ssdtools Shiny app would need to be modified, possibly to include alternative ‘tabs’, e.g. a ‘tab’ for ‘research mode’ (perhaps similar to that already available) that is flexible and allows extensive exploration by the user, and a ‘tab’ that is clearly for use in official GV derivation, where the expected ‘best practice’ method is set by default, with minimum possible flexibility.

3 ACCOUNTING FOR UNCERTAINTY IN GUIDELINE VALUE DERIVATION

There is a need to recognise the inherent ‘uncertainty’ in ecotoxicity data, regardless of the size of the dataset. Variability in biological tests is invariably large, even across repeated tests using the same substance and test organism and/or endpoints. This uncertainty can be attributed to numerous practical, technical, chemical and biological factors, and is further exacerbated when attempting to translate the results of laboratory tests to the natural environment. It is important to be aware that this problem applies to both large and small ecotoxicity datasets, as it is the uncertainty in the underlying included toxicity data that is not being accounted for.

We need ways to capture this inherent uncertainty in SSDs, regardless of whether this represents uncertainty in underlying toxicity values and/or underlying data reliability, uncertainty associated with large responses between tests (repeatability), or other ecological factors.

Incorporating all sources of uncertainty is complex, and the most critical and obvious place to start is with propagation (or capturing) of the uncertainty around the individual toxicity estimates (e.g. EC_x/NEC) from which SSDs are constructed. Various approaches are available for achieving this (Gottschalk and Nowack, 2013; Kon Kam King et al., 2015) and the relative advantages/ disadvantages need to be evaluated. Some methods may be ‘theoretically’ more statistically correct but require a high density of data for the SSD.

One question raised by respondents and discussed in the workshop was: Is it better to incorporate all of the toxicity data into an SSD, to better account for variability? This question was threefold, is there value in: (i) incorporating all toxicity estimates (e.g. EC₁₀, NOEC, NEC, etc.) for all species (when there are multiple equivalent endpoints); (ii) including the error around the individual toxicity estimates? or (iii) incorporating all the concentration-response data into the SSD. What is the overall cost-benefit? Does it affect the estimated PC_x values or just the confidence limits around them?

3.1 Capturing uncertainty in underlying toxicity data

The general consensus at the workshop was that the ‘gold standard’ for SSD derivation would include all of the underlying concentration-response (CR) data. While this should be the standard ‘forward-looking’ methodology, any approach to be adopted in the short to medium term must also allow ‘backward compatibility’. Criteria for both forward and backward compatibility include:

- (i) a necessity to continue to accept historical NOEC data in the short- to mid-term;
- (ii) the ability to use include the 95% CIs of a toxicity estimate (e.g. EC_x/NEC values) if the raw CR data are not available;
- (iii) where more than one toxicity estimate exists for a species for the same (most sensitive) endpoint and under the same testing conditions, the ability to include all of the estimates in the derivation rather than using the geometric mean of the estimates (see Section 3.2; and
- (iv) where available, to use the raw CR data.

There are Bayesian mixed modelling statistical methods developed for deriving SSDs using all underlying SSD data, as well as using the interval estimates from the EC_x/NEC values. Proctor (2018) compared four approaches to deriving SSDs, including using: (i) all CR data, (ii) interval estimates, (iii) single point estimates, or (iv) a mix of interval data and single-point estimates for SSD derivation using simulated data. It was shown that the use of individual uncertainty increased the confidence interval around PC_x values, without greatly changing the PC_x value. This was also seen in the work by (Zhao and Zhang, 2017). However, Proctor (2018) argued that the widened confidence intervals increase the likelihood the true PC_x for the population is contained.

Bootstrapping and Monte Carlo simulation methods may provide an alternative means of propagating uncertainty in underlying endpoints into the SSD. One approach proposed during the workshop (and identified for further development and evaluation) is to use a Monte Carlo simulation based on the underlying uncertainty in the NEC/EC_x values, and then base inference on many thousands of fitted SSD curves. Such a Monte Carlo simulation could be relatively simply built using either the posterior probability density samples of each endpoint value (in the case where underlying EC_x/NEC was estimated using Bayesian fitting procedures), or a sample drawn from the best-fit statistical distribution to the EC_x/NEC point estimate and supplied confidence bounds or combining both where appropriate. The approach would be similar to that adopted by (Gottschalk and Nowack, 2013). The utility of such an approach and the advantages and disadvantages compared to the Bayesian hierarchical modelling approach warrants further investigation, both with small and large datasets.

Regardless of the statistical methods used for capturing and propagating uncertainty in SSD estimation, more effort needs to go into developing user-friendly tools and interfaces, to increase accessibility to those without access to specialist statistical programming skills and the ensure that the methods are accessible to a wide range of practicing ecotoxicologists.

3.2 Incorporating all toxicity estimates for a single endpoint

Repeatability among laboratories using the same suite of species can be highly variable, particularly when the test chemical exhibits high toxicity, high volatility and/or low aqueous solubility. One respondent asked the question: “What does it mean if a test, repeated say three times over several years, produces toxicity estimates an order of magnitude different and how should the data be used in GV derivations?”. This question relates closely to another that arose in the context of multiple experiments: “Should analysts be using an average EC_x from replicate tests or estimating a single EC_x for pooled data from all replicate tests, and are there criteria or a test to justify pooling data?”. Provided that the raw CR data are available from all of the experiments, a single EC_x/NEC value from the combined CR data sets could be estimated (and/or used in the SSD derivation as described above). This should only be done when the experimental conditions are sufficiently similar (e.g. test endpoint, test duration, test water quality) or, where differences occur, these are known to not influence the toxic response. Combining CR data from multiple experiments does increase the statistical complexity of the model fitting, because there may be variations due to the different experiments leading to non-independence of data within each of the experiments. Statistical tests should be performed assessing if there are differences in the relationship the maybe attributed to

the different experiments, and if there are, these should be accounted for in the endpoint estimation. Non-linear mixed effects models (Pinheiro et al., 2013) represent a possible approach to cope with this non-independence in a statistically rigorous way if differences among experiments are evident. Developing guidance and user-friendly/accessible methods for fitting such statistical models to combined CR data from multiple experiments is clearly required to improve their uptake in ecotoxicological statistics.

Where all the raw data are not available, which would be the case most of the time, fitting a pooled CR model is not possible. In this case, a geometric mean of the multiple toxicity estimates is currently recommended, again, assuming the test endpoint and test conditions are the same (Warne et al., 2018). Using a geometric mean potentially masks true uncertainty in the toxicity estimate, potentially leading to overconfidence in the resulting PCx values. An alternative approach that allows the individual uncertainty in the underlying ECx estimates to be propagated, is to use all of the individual toxicity estimates, yet weight them to ensure that their total contribution sums to 1 (i.e. that all species contribute equally to the SSD). A weighted contribution of multiple endpoints in the SSD estimation could be easily achieved using the same Monte Carlo simulation framework described in Section 3.1 above, by weighting the proportional contribution that each toxicity estimate distribution has in each iteration of the Monte Carlo simulation. Note that such a weighting scheme is simply focused on allowing multiple equivalent endpoints to be included in the SSD, with their associated uncertainty, with all available treated equally and a total combined weight equivalent to a single species endpoint value. Issues around weighting of endpoints based on data quality and other criteria (such as taxonomic independence) are discussed in more detail below.

3.3 Dealing with other sources of uncertainty

A range of other issues associated with uncertainty were raised by survey respondents, not all of which were discussed at the workshop, or can be easily and clearly resolved. They include:

- How to propagate uncertainty of endpoints converted between acute and chronic values, or that are derived through means other than CR relationships? There was agreement that the ideal situation was to not use acute data at all (unless the contaminant issue was more relevant to acute exposures/toxicity), and it was conceded that the current guidance around when to use acute data and how to treat it was about as good as we could manage.
- How do we account for uncertainty associated with specifically included toxicity values (species) and outliers, and should we? This is particularly a problem when there are outlying values in a dataset (primarily at one of the tails of the distribution) that have a large influence on the resulting GVs. This is always more of an issue for small datasets. True outliers can be safely removed, however, ‘aberrant’ (but nonetheless legitimate observations) should remain in the analysis and will contribute to the overall uncertainty estimate. There is a wide range of cross-validation approaches that are routinely used in other areas of statistics, albeit in areas such as ecological spatial modelling which tend to be data rich. While cross-validation methods adopting “testing versus training” strategies may not be applicable in the typically data-poor ecotoxicology setting, others such as ‘leave-one-out’ (LOO) may be appropriate. While not explored further at the workshop, the utility of adopting such approaches for capturing uncertainty around data-inclusion should be

certainly be a target area for future investigation. The possibility of bootstrapping data to overcome uncertainty in the data points was suggested, as you can sample the ranges of uncertainty no matter the model type used.

- How is uncertainty associated with the fitted model captured? Is this resolved by model-averaging? Uncertainty in the fitted model is an inherent feature of statistical modelling, and all current procedures account for uncertainty in the estimated PC. The model-averaging approach discussed above allows a more flexible approach to dealing with the different fits arising from the choice of different models. Given the potential advantages associated with the use of model-averaging and mixture-modelling in SSDs, consideration needs to be given to how to integrate methods for propagating uncertainty using these procedures, so that the underlying data uncertainty, as well as model uncertainty can both be effectively dealt with.

3.4 How should uncertainty and variability be used?

The uncertainties of deriving PC95 and PC99 from NEC, NOEC and EC5-EC10 data can be very large. Many of the ideas and suggestions discussed above are likely to increase estimates of uncertainty, rather than having any large impact on the estimated PCx values, although this still needs to be evaluated, once some of the suggested methods have been explored. Currently, the confidence limits associated with GVs do not form part of the formally reported GV documentation. A reluctance to report on and/or discuss uncertainty is in part a consequence of the fact that SSD modelling is, by necessity, often undertaken using inadequate sample sizes, which can cause instability and lack of robustness – particularly in the calculation of confidence intervals. However, until an alternative approach presents itself SSD modelling remains the ‘best’ we have.

The proposed improvements to the approaches used for modelling SSDs discussed above (such as model averaging and propagation of underlying endpoint uncertainty) will hopefully improve the robustness of CI estimation and increase confidence in their use. Where sample sizes are sufficient and we have robust CI estimation, however, the question remains as to how a regulator should use the 95% CI associated with published GVs and/or what national advice is provided on what the CIs mean or how they should be interpreted. Ideally, uncertainty in GVs should be reported and used, for several reasons:

- (i) to ensure protection (e.g. safety factors or lower confidence bounds on best estimates);
- (ii) for transparency concerning confidence; and
- (iii) for extrapolation from incomplete data. However, the CIs need to be able to play a clear and meaningful role in regulatory and general water quality management processes.

Without appropriate guidance and a clear purpose for their use, there is a risk that the upper and lower CIs of a GV will be selectively used by interest groups to justify their interests (i.e. industry may advocate for an upper 95% CI, whereas an environmental lobby group may insist on a lower 5% CI value). Within a formal decision science framework, the uncertainty in the underlying GV estimate (and associated CI values) can actually be imbedded into the decision context, where the actual GV value used is selected based on the relative cost of causing environmental harm (see Fisher et al., 2018).

Confidence limits around PCx values derived from SSDs can also be used to compare different values, for example, if a researcher wants to compare one PCx to another (e.g. one that corrects for bioavailability and one that doesn't). Ideally, statistical comparisons between different PCx values should be made using models fitted to the full SSD datasets (see Section 2.7), not just the resulting PCx values. However, provided that the uncertainty has been properly accounted for, robustly estimated confidence bounds do provide a means of statistically comparing PCx values. Even if a more rigorous statistical treatment of uncertainty results in larger confidence bounds, such statistical rigour is essential if these bounds are to reflect the true uncertainty associated with derived GVs. If rigorously estimated confidence bounds are very wide, this provides important information that there is uncertainty in the level of protection actually provided by the GV, and other weight of evidence or precautionary approaches may be warranted. Consistent with this, ANZG (2018) recommends GVs be used as one line of evidence in a weight of evidence approach. Clearly, more guidance on how to deal with the confidence limits is needed. Such guidance should include how confidence limits and uncertainty can be incorporated into management decisions through formal decision science frameworks (for example, see Fisher et al., 2018), as well as providing guidance on wording around reporting to ensure confidence in the GVs and their subsequent regulatory use are maintained, despite the acknowledgement of (real) uncertainty. There is an opportunity here to advance this discussion and explore the possibility of a transition to a different use of GVs and their associated confidence limits within the regulatory and general water quality management context.

3.5 Recommendations for capturing and reporting uncertainty in GV derivation

- There is a range of existing methods for capturing uncertainty in underlying toxicity CR data that need to be thoroughly explored, including Monte Carlo simulation and Bayesian mixed modelling approaches. Possible methods need to be reviewed, and where appropriate compared, using real and/or simulated data.
- Methods for allowing multiple equivalent endpoints to be included in the SSD need to be developed, so that their underlying uncertainty can also be taken into account.
- While not explored further at the workshop, the utility of adopting approaches such as leave-one-out cross validation for capturing uncertainty around data-inclusion should be a target area for future investigation.
- Regardless of the statistical methods used for capturing and propagating uncertainty in SSD estimation, more effort needs to go into developing user-friendly tools and interfaces, to increase accessibility to those without access to specialist statistical programming skills and to ensure the methods are accessible to a wide range of practicing ecotoxicologists.
- More guidance on how to deal with the uncertainty and confidence limits is needed. Such guidance should include: (i) how confidence limits and uncertainty can be incorporated into management decisions through formal decision science (see Section 7.1), and (ii) wording around reporting to ensure confidence in the GVs and their subsequent regulatory use are maintained, despite the acknowledgement of (real) uncertainty.

4 INPUT TOXICITY DATA FOR SSDs

4.1 Toxicity estimates

One of the key issues raised by numerous respondents related to the use of different toxicity estimates from the concentration-response data, such as NEC versus ECx. The current guidance indicates that: *“The preferred order of statistical estimates of chronic toxicity to calculate default and site-specific GVs is: chronic NEC, EC/IC/LCx where $x \leq 10$, BEC10, EC/IC/LC15–20, and NOEC. While all of these acceptable statistical estimates of toxicity are not numerically the same, they are all treated as equivalent for the purposes of deriving GVs”* (Warne et al., 2018).

Confusion around this statement arises due to the final statement *“they are all treated as equivalent for the purposes of deriving GVs”*. Use of the NEC is clearly the preferred estimate of toxicity for GV derivation, although ECx values (e.g. EC10, EC50) are becoming the dominant toxicity estimates reported in the literature (Warne et al., 2018). The fact that all the toxicity estimates within the above hierarchy are considered ‘acceptable’ stems from the recognition that NEC values are hard to estimate (see below, deriving NEC) and that there are very few toxicants for which they are available, making GV derivation based solely on NEC estimates currently impossible. The intent of the current guidance is such that where an NEC estimate is available, this should be used in the derivation, unless there are limitations associated with the dataset that render the data of unacceptable quality. That all the listed toxicity estimates are treated as “equivalent” is a direct result of the complexity that arises when combining multiple toxicity estimates across species in the GV derivation process (see Section 7 below) were this assumption not made. To not treat these estimates “equivalently” requires the development of a weighting scheme, which introduces an array of issues that cannot be obviously, consistently and rigorously resolved (see Section 5).

There is considerable uncertainty and inconsistency in the way individual ecotoxicological datasets are quantitatively analysed to derive toxicity estimates. This includes differences in the statistical packages and modelling methods used, which can produce slightly different results, differences in the toxicity estimates derived, as well as consideration of how to deal with the case when there are multiple tests. We need straightforward robust guidelines on how best to estimate endpoints and accessible and easy to use tools to support accepted best methods.

4.1.1 Estimating the no effect concentration (NEC)

The NEC is the preferred endpoint for use in GV derivation because it represents the concentration at which there is no impact of the toxicant on a given species. Note that a NEC is distinct from a NOEC which simply represents the No-Observed-Effect-Concentration that is sometimes reported when the experimental design was insufficient to parametrically model the CR relationship (e.g. when there are too few experimental treatment levels) (Fox and Landis, 2016). Through the SSD approach, the goal is to estimate a PCx value that represents the concentration for which no more than x % of all species will be affected (or conversely, >x % will not be affected). Since ‘affected’ versus ‘not affected’ is a binary outcome, the input data for the SSD must, to be consistent, represent no effect thresholds (i.e. NEC values). Any ICx, ECx, or LCx estimate represents an effect of some magnitude

and therefore cannot technically be a surrogate for a NEC. However, until there are sufficient no effects data (i.e. NECs), the practice of mixing no effect and various low effect (e.g. EC5, EC10) data remains the most pragmatic way to increase sample size and also to reduce the reliance on the NOEC for use in SSDs. However, a PCx estimated from such an SSD can only be considered to represent the concentration at which at least x% of species will experience no *adverse* effects, as opposed to no effects at all (assuming that the low effect estimates that are used actually represent effect sizes that will not adversely affect an organism's ability to contribute to the ongoing sustainability of the population).

While NEC values are clearly more consistent with the objective of SSD modelling, they are only rarely reported in the literature. Lack of use of NEC in the current ecotoxicological literature is likely due to perceived inaccessibility of statistical methods for deriving the NEC, as well as the oftentimes poor experimental design precluding its estimation. Also, it is currently not recommended as a standard toxicity estimate in the various toxicity test methods and associated statistical analysis guidance documents that exist (e.g. OECD, 2015; ECCC, 2018, as two recent examples), and which drive much of the toxicity data and associated toxicity estimates reported in the literature.

The statistically appropriate way of estimating NEC is through a 'segmented' regression (also known as 'broken stick' 'split-point', and 'piece-wise' regression) approach. In this approach, some mean value of response is modelled across a range of concentrations for which no effect occurs, with some kind of decay function used to model the change in response following an estimated transition point, which directly corresponds to the NEC value. There are several published methods available for estimating NEC, including a Bayesian method implemented in the package *winbugs* (Fox, 2010) based on the method of Pires et al. (2002), that can now also be implemented in the popular and freely available statistical programming language 'R' through the package 'R2jags' (Su and Yajima, 2015) among others. While there is sometimes the perception that Bayesian methods "are difficult to use because of the necessary subjective probability distributions", this is generally not a problem in practice. A range of uninformative priors are available that largely eliminate the problem of subjectivity associated with Bayesian inference, and the posterior probability sample that is obtained as a by-product of the Bayesian MCMC fitting method represents a rich source of information around parameter uncertainty unmatched by frequentist approaches. Even so, frequentist approaches for estimating NEC are also readily available through the package 'drc', also in 'R'. While software to implement both methods is also freely available, use of both packages requires reasonably sophisticated programming skills, and such specialist skills are not always available to practicing ecotoxicologists. Friendlier user interfaces for applying such statistical methods, perhaps through the use of 'Shiny' from 'RStudio' (see discussion above on SSD model averaging and *ssdtools*), may pave the way to greater uptake of NEC by a broader cross-section of the ecotoxicological community.

Aside from the logistics around statistical computation of NEC, the ability to fit the theoretically appropriate statistical models are often hindered by the quality of the underlying concentration-response data. Fitting a segmented regression model (as implemented in Fox, 2010) requires sufficient treatment concentration values both before and after the NEC transition point being estimated, such that the 'mean' value of response below the transition point as well as the decay

function can both be estimated reliably. Often concentration-response data lack either the necessary density of treatment concentrations, and/or have treatment concentrations that are inadequately distributed with respect to the important transition points to be properly estimated.

Ensuring the concentration-response data are sufficient for estimating an NEC requires prior knowledge of a likely transition concentration value. This may be achieved through a review of toxicity concentrations for similar species and/or toxicants or, where little information is available, through a range-finder experiment. Range-finder experiments are an extremely important and efficient means of identifying the effects range upon which to focus more comprehensive definitive CR experiments and should always be carried out where possible. Also, running more than one definitive test, with subsequent test concentration ranges being informed by the previous tests, will always help with the definition of the CR relationship. However, clear guidance on the best experimental design for NEC derivation is urgently needed if we are to increase the use of NEC in water quality GV derivation.

Even for a well-designed concentration-response experiment, it can sometimes be impossible to estimate a reliable NEC value. This occurs for example, when there is no evidence of a 'threshold', which can happen when the response simply changes as a gradual function of concentration, or when the compound is toxic at any concentration. In this case, an EC10 is generally considered a good/acceptable endpoint.

Clear advice on some of these issues does exist and is available in the published literature and/or specialist ecotoxicological statistics courses. However, the key issue is that this type of advice needs to find its way into standard toxicity test methods and ecotoxicology guidance documents published by, e.g. OECD, USEPA, ISO and Environment Canada. For example, a very recent revision of an Environment Canada *Hyalella azteca* toxicity test method, still recommends 5 replicates per treatment (Environment Canada, 2017). While it does at least recommend a minimum of 7 and preferably more treatments, the overall experimental design guidance is still inconsistent with preferred statistical practice (note that this is an example and is not attempting to single-out a specific protocol or jurisdiction). This applies to whether one is looking to estimate a NEC or an ECx. Broadscale change won't be achieved until relevant test methods and guidance documents are updated to include known best practice. This issue is similar to that of the NOEC, the use of which is now being phased as a result of recommendations against its use have been included in guidance.

4.1.2 General considerations

Aside from the discussion around estimating NEC values above, there is a clear need for more guidance on, not just the design of future CR experiments, but also how to best use the data that do exist. OECD (2006) provided comprehensive guidance on how to analyse CR data, including recognition of the NEC as a measure of toxicity that can be estimated, but did not provide detailed guidance on the NEC. The international ecotoxicology community has been slow to evolve its statistical practices, and more effort is required to effect necessary change (van Dam et al., 2012). Fox (2010, 2018) described the NEC and its estimation in greater detail, including specific considerations for experimental design. However, the information on NEC and associated

experimental design considerations still need to find its way into more formal guidance documents in order that adoption is more successful.

In addition to the segmented regression above, what models are available for fitting and calculating concentration-response data to estimate EC_x? Even more than for NEC, there is a wide range of statistical packages available, some with intuitive user-friendly graphical user interfaces (GUIs) (e.g. ToxCalc, CETIS, SigmaPlot) although many of these lack the flexibility offered by the open source platform, R. Regardless, with the wide range of approaches and models that can be used for fitting CR data, clearly some guidance around ‘best practice’ methods is potentially warranted. The same data fitted using different approaches can yield different EC_x values, reducing confidence in GV derivation results.

Some general considerations include:

- (i) Bayesian or frequentist? Each have their advantages and disadvantages;
- (ii) Are model averaging and mixture modelling approaches potentially useful here? (see discussion on SSD model averaging and ssdtools);
- (iii) Are there minimum levels of concentration required to use segmented regression methods for NEC below which EC₁₀ estimation is required?
- (iv) What is the best approach when treatment levels are low or insufficiently dense at the low response end? and
- (v) Does “% of control” in CR relationships matter? Should we be normalising to controls?

4.2 Acute versus chronic endpoints

The use of converted ‘acute’ data in the derivation of ‘chronic’ GVs is frequently debated. It was generally agreed by the workshop attendees that acute data (converted to chronic “equivalent” data) should not be used in GV derivation. However, given the paucity of chronic data for many toxicants, it was also generally agreed that it is of more value to use, rather than ignore, such data, but to do so with caution and transparency.

Currently, the formal GV derivation method (Warne et al., 2018) recommends that only chronic data be used if: (i) there are five or more chronic values, or (ii) if the fit of the SSD model to a larger set of chronic values is poor (noting, however, that there are instances where an acute GV might be required using only acute data). However, this is often not straightforward. Taking the former case, a degree of professional judgment is often required if, for example, the chronic dataset consists of five or six values while the acute dataset consists of 20 or more values, or even where the size of a small chronic dataset can be doubled by the inclusion of (converted) acute data. In the case of the second situation above, determining the adequacy of fit of an SSD, and whether the fit of one (chronic only) SSD is better than another SSD for a different (chronic + converted acute) dataset, will always require a degree of professional judgment. There are also issues around the use of default acute to chronic ratios (ACRs) vs experimentally-derived ACRs to convert acute data to chronic “equivalents”. Although the latter are considered more appropriate, the question remains, how reliable are they, and how is the uncertainty associated with the ACR captured in the overall SSD derivation? The potential unreliability of experimentally derived ACRs stems in part from their being generated on

single species, yet they may vary by orders of magnitude between different species and yet a single value is applied to all species. The need to use either the default or experimentally-based ACRs is one reason for the preference to use chronic toxicity data over acute data.

The Australian and New Zealand default GVs (DGVs) aim to protect species from life-long exposure to toxicants. However, organisms and ecosystems are frequently exposed to toxicants for relatively short-term periods, e.g. from a spill, deliberate releases under severe weather conditions, or intermittent releases of pesticides from agricultural land. The DGVs are not appropriate for such short-term exposures, being too conservative, and there have been calls to derive GVs based on acute toxicity data, similar to those that the USEPA has always had, and those used in Canada.

The derivation of short-term exposure GVs does have merit and would not be a particularly large undertaking as the data that have already been collated when deriving DGVs could be used. Any such GVs should be derived using the methods set out in Warne et al. (2018) with the exception that acute toxicity data would preferably be used. Such short-term GVs are essentially a form of site-specific GV and their derivation should be transparent and fully justified. Consideration would also need to be given to how to deal with how large and how frequent exceedances could be before further action is triggered.

Aside from the debate around the use of 'acute' toxicity estimates in GV derivation, one respondent also raised an issue around assigning acute and chronic data only on the basis of the duration of the test, rather than on the relevance of the endpoints to the protection goals (e.g. acute toxicity, population sustainability, ecosystem function etc.). This suggests that more discussion, guidance and/or 'rules' may be warranted around the definition of 'acute' versus 'chronic' endpoints across a wider range of species and response variables.

The use of theoretically derived NECs has been suggested as an alternate to acute data, which may overcome the need for acute-to-chronic ratios, as it provides a theoretically time-independent measure of toxicity. This can be calculated with the complex toxicokinetic-toxicodynamic methods (Baas et al. 2010; Jager et al. 2011), often recommended in the European guidelines (OECD 2006). More directly, NECs can also be estimated with the segmented regression model, originally established for ecotoxicology by Pires et al. (2002) and established for use by Fox (2010). The use of NECs may also overcome the difficulties of comparing individual toxicity estimates and SSDs with tropical and temperate data. They can also be used to reassess existing acute data, limiting the use of LCx estimates (Proctor, 2018).

4.3 Other temporal considerations

Currently, water quality GV derivation is based on chronic exposure data. Many respondents had concerns around the fact that testing durations do not necessarily reflect the conditions that species will actually experience. Testing timeframes ought to reflect the likely exposure timeframes and concentration/response profiles actually likely to be experienced in the field. While this should certainly be the goal of site-specific GVs, broad recommendations cannot be made for default GVs, because exposure profiles vary so broadly.

It seems intuitive that time-dependent GVs would offer more realistic protection and a greater level of management applicability. However, the amount of experimental data that would be required to derive these for multiple toxicants would prohibit their routine development. Furthermore, while we try to make toxicant concentrations constant in laboratory chronic toxicity tests (over duration of hours to weeks depending on the species), in the environment they are often temporally highly dynamic.

In principle, GVs need to define the magnitude, duration and frequency of exposure for protecting aquatic life from acute and chronic effects of a contaminant. Often, derivation of GVs focuses only on magnitude and overlooks the attributes of duration and frequency of exposure. In addition, few exposure studies consider recovery following the cessation of exposure to the toxicant.

Incorporating all of these temporal issues is not trivial, and will not be easily resolved, except for given site-specific GV derivations (and even then, the issues remain challenging).

One issue that could potentially be tackled is the fact that we currently have no short-term (acute) GVs. We need an approach that addresses the impact of acute toxicant concentrations and time of exposure (time-averaged concentrations). The status quo of applying chronic GVs is conservative and therefore appropriately precautionary. In some instances, however, acute (short-term) GVs may be warranted, for example in the case of one-off or intermittent short-term spill scenarios. If we are to work towards including acute GVs then we need to be very prescriptive about when it is acceptable to use them in risk assessments or EISs, e.g. for single or infrequent intermittent exposures and not where frequent repeated spiked exposures are expected and may lead to cumulative impacts. The derivation and implementation of GVs based on short-term pulse exposures has been well demonstrated for magnesium associated with a mine water discharge (Hogan et al., 2013; Sinclair et al., 2014). The USEPA (and Canada and the EU) has always had both short and long-term water quality standards and have specified the frequency of exceedances that are permitted. Such an approach should be examined for its relevance to Australian situations.

4.4 Reliability of ecotoxicity data and use of poor-quality ecotoxicity data

There are fundamental questions about the data we are dealing with, their reliability and appropriateness for deriving protective concentrations. There is a need to deal with reliability of all data that are incorporated? Included? into the SSD. These issues go beyond some of the philosophical and statistical details discussed above and include factors associated with the experimental conditions themselves. Examples include: issues associated with particulate vs dissolved contaminant toxicity at high concentrations; feeding vs non-feeding of test organisms; and flow-through vs static tests; where they occur, how do we deal with decaying concentrations in laboratory tests, where similar losses might not occur in the field? There is a myriad of factors that can influence the outcome of a concentration-response experiment (see also Appendix B, 'Exposure mechanics/characteristics').

There is a well-established system for scoring the quality of underlying toxicity data for acceptability for deriving GVs in Australia and New Zealand (Warne et al., 2018). The data quality assessment process scores toxicity data against numerous aspects, mostly related to whether a test was performed and recorded well. A quality score of 50% or more qualifies the data as being acceptable for inclusion in an SSD dataset. Thus, a few flaws relating to various aspects of the method and/or results will not necessarily render it unacceptable for inclusion in GV derivation. Moreover, not all experimental flaws can be picked up by the quality assessment system, and it is often left to professional judgment as to whether specific flaws invalidate a test or warrant its exclusion from GV derivation.

Despite quite good guidance on inclusion criteria and quality scoring, a range of respondents raised issues around the use of some endpoint data (Table 3). These issues included aspects of ‘relevance’, such as what is considered a true adverse effect endpoint (i.e. sub-lethal changes that are of ambiguous population level relevance); or the issue of geographic distribution (e.g. tropical versus polar, geographic region such as US versus Asia). For the most part, guidance on these issues in Australia is clear. Endpoints should only be used where they have been shown to have population level relevance to survival, growth or reproduction. This overcomes the fourth issue raised in Table 3. Furthermore, there are rarely enough reliable data that are directly relevant to Australia to be able to use them in isolation of data that are not as relevant (e.g. northern hemisphere species), thus all reliable data are used. Even where there may be sufficient data to consider, e.g. geographic separation, it is generally preferable to combine all data to maximise biodiversity and data quantity going into the SSD (Peters et al., 2019). Care should be taken when comparing results across geographic regions and/or climatic factors as differences may occur due to other experimental factors, such as experimental water quality conditions, rather than being true geographic differences (Mooney et al., 2019).

Table 3: Key issues identified by respondents relating to the use of poor quality data and/or relevance of endpoints

The exclusion of otherwise reliable and relevant data based on supposed, assumed, or unevicenced differences in sensitivity between species within taxonomic groups, based upon, e.g. geographical distribution, temperature, etc.
The inconsistency between approaches/ methods applied to assess the reliability and relevance of data between different jurisdictions.
The need for greater frequency of rejecting toxicity values (e.g. EC10 values) extrapolated from relatively poor concentration-response curves.
An over-reliance on studies / endpoints of questionable reliability / population level relevance. Examples include endpoints such as up or down regulation of genes, enzyme induction or histological changes that cannot clearly be translated into effects on population survival, growth or reproduction should not be used in guideline derivation.

The fundamental trade-off that such questions represent can be summarised as: “Is it better to have more points and include low-quality data, or just stick with fewer high-quality data points?” According to one respondent, there is a view that there is a general preference to accept rather than reject data. Although the quality assessment process provides an objective means for rejecting data and, in fact does result in many data being rejected, the process still has subjective elements, and it

is foreseeable that there may be a tendency to want to retain data in the face of very small datasets. However, it is clear in the Australian method of deriving GVs (Warne et al., 2018) that only acceptable quality should be used and if there are insufficient to use the SSD method then appropriate quality data for another media (e.g. fresh or marine) or different exposure durations (e.g. acute or chronic) can be used. It was also suggested that existing test accreditation processes do not ensure the production of good quality data and, in fact, can be counter-productive due to them being seen as a stamp of approval. It is important to recognise that accreditation processes do not capture all the elements of toxicity testing that help ensure the production of reliable and relevant data, especially the latter. Also, accreditation focuses on formally approved test methods that are not readily adaptable to advances in experimental and statistical practices. The data quality systems used in Europe in particular, place great importance on the use of standardised methods to generate toxicity data. This is not the approach adopted in Australia and each study is assessed on its merits.

Toxicity data that appear as outliers may make their way into datasets used to derive GVs even if generated using non-standardised test protocols with no replication of study findings. As already stated above, true outliers (that represent erroneous or invalid data) should be removed. However, 'aberrant' observations should remain in the analysis and will contribute to the overall uncertainty estimate. If the value of aberrant endpoints is very low, they can have a major influence on the resulting GV, which further emphasises the need for close scrutiny of toxicity data, and the necessity for the rigorous inclusion criteria currently adopted within Australia. The quality assessment and data-screening processes should be frequently re-examined and modified if appropriate.

A significant problem with allowing the use of low-quality data (as a means of increasing dataset sample size) is that it can set a precedence for not having to generate the good-quality data that we actually need. Masking of the need for better quality data is further exacerbated when the uncertainty associated with poor-quality data is not propagated through into the confidence bounds associated with estimated PCx values, and because the level of confidence in the GVs is rarely reported.

Overall, the workshop group felt that criteria around data quality are currently well established and that the use of poor-quality data that fail these criteria should not be considered. Ultimately, where there are few data available, alternative methods to SSD should be considered, rather than inclusion of inappropriate data.

4.5 Species selection

Another broad overarching topic related to the input toxicity data are issues around which species to include, a topic raised repeatedly in the key issues survey (Table 4). The statistical legitimacy of the SSD paradigm rests on the assumption that the species selected represent a random sample from the larger population of all species in the ecosystem being protected (Fox, 2015). This assumption is rarely met, although there are selection bias correction factors that can be used to accommodate deviation from this assumption to some extent (Fox, 2015).

Aside from the fundamental assumption of random selection, consideration needs to be given to the relevance of the ecotoxicity dataset to the ecosystems to which they are being applied. This is a fundamental requirement for site-specific GVs but also potentially needs to be given consideration for default GVs. Are existing datasets sufficient to ensure that the protection goals are met in the environment? Is there any reliable evidence to support or refute this? For example, insects can make up the bulk of the diversity in some aquatic communities, but rarely make up much of the ecotoxicity dataset because they are difficult to test in the laboratory. Future emphases for species selection could potentially focus on development of test methods for typically under-represented taxa. Other standards focus on a single species, such as drinking water standards which aim to derive safe levels of concentrations for humans. The derivation of a water quality GV aims to derive (safe?) concentrations for ecosystems – this is a fundamentally challenging task which could be perceived as intractable.

4.5.1 Sensitive species or ‘keystone’ species?

The idea that you can protect a known percentage of the population by ensuring that toxicant concentrations remain below a given threshold is fundamentally problematic, because communities live in complex ecosystems. If there are key species that underpin the functioning of that ecosystem that are highly sensitive, then this could lead to a cascade of impact.

Is it critical that the chosen test species are protected? Are less-sensitive species of the same taxonomic group acceptable in maintaining a healthy ecosystem? What defines a healthy ecosystem? Some species might be more important than others (i.e. the impact to the ecosystem can be significant different), but how to determine the key species?

We need standardised approaches and justifications for including specific species, and/or potentially the ability to focus on key taxa, for example “habitat-forming” groups. The ability to weight taxa in the SSD according to their relative importance may alleviate this (see below) but approaches entirely separate from an SSD may be more appropriate. This ability to focus on key species may be most pertinent when data sets and taxonomic coverage is limited and likely not very representative of the diversity in the receiving community.

This issues around species selection, taxonomic representativeness and ecosystem relevance are many faceted, and were beyond the scope of discussions at the workshop. Certainly, there are no clear recommendations for how these issues would be resolved for a national guideline perspective. For site-specific and application of specific guidelines these issues are critical and warrant some thought and investigation.

Table 4: Key issues identified by respondents relating to species selection and taxonomic representativeness

If the SSD framework is to be used, how does the species selection impact GVs?
While there is guidance on how many species are required from separate taxonomic groups, has there been any sensitivity analysis to underpin this?
Should there be weighting (see below) to allow greater weight of less similar species?
Are GVs biased by species selection within a taxonomic group?

Look at key species or taxa that are relevant within the Australian nearshore and deep-water ecosystems, that are wider than the list currently used in standard ecotoxicity testing, notably corals.

4.6 Recommendations regarding input toxicity data

We need:

- Straightforward robust guidelines on how best to estimate endpoints and accessible and easy to use tools to support accepted best methods for fitting CR datasets.
- Friendlier user interfaces for applying methods for calculating NEC and/or highly robust ECx values (perhaps developed through 'Shiny')
- Guidance on designing ecotoxicological experiments that allow calculation of NEC and/or highly robust ECx values.
- More discussion, guidance and/or 'rules' around the definition of 'acute' versus 'chronic' endpoints across a wider range of species and response variables.
- Future emphases for species selection could potentially focus on development of test methods for typically under-represented taxa.

5 WEIGHTING OF INDIVIDUAL DATA POINTS / DATASETS

Several respondents raised issues around the concept of ‘weighting’ the individual data points in the SSD, for example: “At the moment NOECs are given equal weighting as EC/IC10 data. This is unfortunate because NOECs have long been recognised as poor estimates of the lowest concentration to cause an effect”; and “Inclusion of weighting of individual data points in an SSD (especially with small data sets) – based on certainty of value, ecological importance, single vs community response, rapid vs conventional testing.” The topic of weighting data points in the SSD was ranked to be of relatively high importance by the workshop participants, however, the workshop group was unable to delve into this in detail. It was highlighted that this topic has received considerable debate historically (Forbes and Forbes, 1993; Forbes and Calow, 2002). At present there is little appetite for including weighting in the SSD for the development of official GVs, although it may be valid for site-specific GV derivation, where valid arguments can be made.

5.1 Weighting based on data quality

While there is a well-established system for scoring the quality of underlying toxicity data (see Section 4.4), these scores are not currently used to weight the data’s contribution to the SSD. Some respondents raised the question: “Should this be done? If so, are there already methods for doing this (within an SSD context, or otherwise)?” Methods for weighting individual data points in the SSD based on their quality score are available and could certainly be included in future platforms developed for SSD fitting. These methods work in the same way that multiple single-species endpoints can all be included and weighted to ensure they contribute the equivalent of a single species (see above ‘Incorporating all toxicity estimates for a single endpoint’). Weighting the SSD data points based on their quality scores has the advantage of avoiding the need for relatively arbitrary cut-offs around data acceptability and inclusion. However, some argue that weighting points by reliability may cause bias in terms of the contributions of the species for which higher quality data exist (i.e. were based on very robust experimental methods), thus making the GV no longer truly representative of ‘x’ species protection by causing further bias towards a few dominant species. This would need to be tested. This potential issue could be overcome if the weighting was applied separately to data for each species and not to the single value representing each species.

5.2 Weighting based on taxonomic representation

Forbes and co-workers (Forbes and Forbes, 1993; Forbes and Calow, 2002) made the point that only a fraction of the species going into the SSD determines the effects threshold. With all species being weighted equally, the loss of any species is considered to be of equal importance to the ecosystem, and keystone, foundation or other important species are assumed to be randomly distributed in the SSD. Batley et al. (2018) concluded that for Australia and New Zealand, “Implicit in the statistical distribution-fitting exercise is the untested, but surely false assumption that the data represents a

truly random sample from the population of all species”. How we accommodate our ‘selection-bias’ in SSD-fitting, is another unresolved issue that requires further research. Weighting might be usefully considered for site-specific assessments and the calculation of site-specific GVs where the information needed to determine the weightings may be available, but as with all deviations from the national approach, the method must be transparent and scientifically valid.”

5.3 Recommendations

Although the current view is that there is insufficient basis upon which to defensibly adopt the use of weighting approaches for SSDs (Batley et al., 2018), more effort in this area might be warranted, including an assessment of any progress made over the past decade. Weighting for site- and application-specific GVs may be appropriate, and investigation around tools that allow this is warranted.

6 SMALL DATASETS

All of the issues discussed above (for example, ‘Reliability of data and use of poor-quality ecotoxicity data’) are compounded in the presence of small sample sizes. What are the best-practice approaches for deriving GVs when sample sizes are small? Should approaches alternative to SSDs be considered? What to do when such numbers are not available? There is a compromise between data availability and the need for guidance of some kind, and more guidance is required around the best practice approach to providing a GV when sample sizes are small.

There was a pragmatic decision in 2000 to permit the derivation of GVs using toxicity from as few as 5 species. However, it is recognised that this severely compromises the statistical validity of the GVs. Although the only other accepted method for deriving GVs from small sample sizes, known as the assessment factor method, typically results in even more uncertain GVs, it doesn’t necessarily follow that SSDs are the best alternative tool to use. At this level of data density, the addition of even a single data point can dramatically change the GVs. This reduces the credibility of the SSD approach. The current derivation method recognises the need for data for at least 8 and preferably 15 species, although 5 species is still allowable (Warne et al., 2018).

When sample sizes are small, individual values can have a large relative impact. One question raised, was: “Is there a way to deal with this?”. Some of the potential statistical improvements to the SSD model fitting discussed above are likely to substantially improve the reliability of GVs derived from small datasets, even if it is only to generate confidence bounds that reflect the true level of uncertainty associated with these estimates. Incorporating uncertainty in the underlying endpoints (see Section 3.1), should ensure that highly uncertain endpoints have less ‘sway’ over the model fit. Similarly, model averaging should help to ensure that uncertainty in the form of the SSD model fit associated with few data is also captured. The adoption of approaches such as leave-one-out cross validation (see above) to account for uncertainty around data inclusion may also be valuable in the case of small sample sizes and should be explored in this context. More testing of any newly developed tools in the specific case of small sample sizes is essential to confirm that these actually improve the quality of inference that can be gained.

There is no doubt that, regardless of the statistical methods used, at low data densities the CIs are likely to be very wide. Historically, these CIs are not provided in the guideline documents and not used to assess GV reliability. However, when rigorously calculated, large CIs indicate high levels of uncertainty in the GV, and this should inform the level of GV reliability. As already discussed, more guidance on the use and interpretation of CIs is warranted and this needs to explicitly include consideration of small datasets. Presenting confidence limits provides a formal way of “Acknowledging the reliability and variability (or lack of reliability) in GVs from small data sets (e.g. 8 or less); specifically, the reliability of GVs for ‘pristine/high-value’ ecosystems (PC99) from small data sets and their application in management decisions.”

In the case of small sample sizes, the question of the ability to include ancillary data becomes more critical. To some extent, this overlaps with the discussion above around the reliability and use of

poor quality ecotoxicity data. The general view was that it is not better to have more points if it means including low quality data. For ecosystems that require high levels of protection (e.g. PC99), the question that needs to be asked is: “Are these values on the extreme tail of the distribution even valid for small sample sizes?” Very few data are being used to estimate GVs on the extreme tail of the SSD, and it seems likely that alternative methods (see Section 7) may yield GV estimates with more inherent management value, despite not being derived to the same strict criteria as the Australian and New Zealand DGVs.

Strict cut-offs around the minimum required toxicity values to use the SSD method, and therefore comply with the Australian guidelines, has the potential to increase the tendency to include poor quality data into the SSD. Research around the trade-offs between more rigorous guidance is needed to decide at what stage we should say the predictions are too poor to be using SSD models, which increases the importance of finding more appropriate alternative methods for small sample sizes (see Section 5). It is worth noting that a GV reliability classification scheme is in place, for which the sample size is one of three factors that determine GV reliability. Thus, GVs based on smaller sample sizes but all other factors being constant will be assigned lower reliability.

6.1 Recommendations

Explore a range of statistical approaches specifically in terms of their ability to cope with small sample sizes, including: model averaging, mixture modelling, uncertainty propagation, and cross validation.

7 ALTERNATIVES TO SSDs

The SSD approach and associated “percentage of the community protected” concept (PC_x) represents an intuitive and useful framework for water quality GV derivation. However, there are a range of concerns associated with the approach that were raised through the ‘key issues’ survey. Many of the challenges and statistical issues associated with SSDs have already been discussed in detail throughout the preceding sections. Beyond these issues, however, there is a desire for better/more flexible modelling approaches than SSDs, particularly in the case of small sample sizes as well as for ecosystems of high ecological value where 99% species protection (PC₉₉) guidelines are required (see **Error! Reference source not found.**). While a thorough review of available alternative methods to SSDs was beyond the scope of the workshop activities, the group did spend some time discussing a limited set of potential approaches, and the underlying motivation behind these, including approaches focused on probabilistic thresholds, as well as community ordination methods. There is also an increasing interest in population modelling and demographic modelling, which use an individual based modelling approach to quantify the impacts of toxicants on populations and local ecosystems that may more accurately assess toxicant risk (Forbes and Forbes, 1993; Salice and Miller, 2003; Gama-Flores et al., 2005; Forbes et al., 2008; Forbes et al., 2016; Gergs et al., 2016; Schmolke et al., 2017; Thursby et al., 2018). Such methods were beyond the scope of the workshop discussions and are not further elaborated here.

7.1 Probabilistic thresholds

Thresholds should be based on a probabilistic assessment of the degree of impact. As currently formulated through standard application of SSDs, this probabilistic statement is linked to protecting a proportion of the species or community (99%, 95% species protection). While this concept is intuitive and can be (and regularly is) incorporated into formal risk assessment, there are some fundamental issues that potentially limit the capacity to imbed SSD-based thresholds into a formal decision science framework. These issues are two-fold: (i) Can a ‘true’ PC₉₉, or even PC₉₅ actually be estimated? and (ii) How can PC_x values be directly linked to environmental and economic cost?

7.1.1 Can a ‘true’ PC₉₉, or even PC₉₅ actually be estimated?

That the derived values truly represent their purported percentages of protection is unlikely, for a range of reasons. Firstly, even the most data-rich SSDs are based on a small sub-set of species and are dependent on those that are included in the ecotoxicity testing, and how these data are fitted by the functional form of the SSD relationship. A different suite of species and a different functional relationship may yield entirely different PC_x threshold values. While there may be viable solutions (see Sections 2.3 and 3.3), it seems unlikely these problems will be resolved entirely through statistical means, and the issues are further compounded when sample sizes are small. Secondly, there is frequently very little confidence at the bottom end of the SSD (i.e. the CIs are wide) and the PC₉₉, in particular, may bear no relationship to the true concentration that would be protective of a high value ecosystem.

7.1.2 How can PCx values be directly linked to environmental and economic cost?

Formal decision science frameworks advocate that thresholds should be derived such that they balance the competing costs associated with failing to indicate potential for harm, versus the cost associated with unnecessary false alarms (Swets et al., 2000). Ideally a water quality management threshold would be set at a level that ensures there is no harm to the ecosystem if it is not exceeded (i.e. has a high true positive probability = TP, aka statistical power, or 1-type II error), and that if it is exceeded there is high confidence that there would be detrimental impacts (low false positive probability = FP, low type I error). In reality, the true positive probability can only be increased at the cost of also increasing false positive probability. The optimal threshold from a decision science perspective should be derived through combining the empirical relationship between TP and FP error rates with a cost model associated with the actual financial costs of committing these two types of errors, i.e. the financial cost of failing to halt an activity that is in fact doing irreversible harm to the environment versus the cost of halting activities unnecessarily when there is in fact no realised harm. While developing the relevant cost model to capture the relative financial impacts of these types of errors may not be practical in the short term, understanding the statistical 'error' rates of thresholds is critical to embedding ecotoxicology within a decision science framework. Focusing the probabilistic problem solely on the percentage of community protected could still be linked to a cost model associated with the cost of lost species, but this would still fail to capture the potential uncertainty associated with the corresponding thresholds, which would yield the actual observed rates of statistical error. Probabilistic thresholds that capture the statistical error associated with thresholds are essential for imbedding thresholds into decision science approaches.

7.1.3 Probabilistic threshold derivation

A range of existing methods has been developed to calculate probabilistic thresholds that also capture the underlying probability density functions of the individual species toxicity data (e.g. Gottschalk and Nowack 2013). A thorough review and evaluation of all proposed methods was not possible during the workshop and warrants further effort. Ideally, any method(s) considered for broader adoption should be relatively easily applied, robust to the problem of small sample sizes, and yield probabilistic thresholds that can be clearly linked to the decision science framework. In addition, the methods should also allow the same 'backward compatibility' criteria as identified above (see Section 3.1) (i) to continue to accept historical NOEC data in the short- to mid-term; (ii) the ability to use include the 95% CIs of a toxicity estimate (e.g. ECx/NEC values) if the raw CR data are not available; (iii) where more than one toxicity estimate exists for a species for the same (most sensitive) endpoint and under the same testing conditions, the ability to include all of the estimates in the derivation rather than using the geometric mean of the estimates (see section **Error! Reference source not found.**); and (iv) where available, to use the raw CR data.

A promising method put forward at the workshop is to simply use an aggregation of the probability density functions for estimated individual species toxicity estimates to derive thresholds that represent known probabilities of community protection. Where no formal probability function is assumed to extrapolate the percentage of total community species protection, such a method is analogous to the bootstrap methods of Newman et al. (2002) but instead combined with a Monte Carlo simulation approach that takes into account the underlying probability density function of each toxicity estimate of the individual species. The proposed probabilistic approach has already

been used by Gottschalk and Nowack (2013) and termed the ‘probabilistic SSD’ (PSSD), which they further imbedded within a formal risk estimation framework. Monte Carlo-based methods readily allow various weighting schemes to be applied, for example, based on taxonomic relatedness, multiple experiments of the same species, as well as weights based on underlying data quality.

Weighting, while perhaps not appropriate for national guideline derivation purposes, may be valuable for imbedding site-specific derivation within a formal decision science framework for deriving useful operational management thresholds.

The described method provides probabilistic thresholds with meaningful error rates, albeit given the available included data. More research needs to be undertaken to assess the relative degree to which the method is dependent on the underlying data, and how robust the approach is given the often small sample sizes available for thresholds derivation.

Where sample sizes are extremely small, and/or when the goal is protection of particular high-value ecosystems or ecosystem functional processes, the use of individual concentration-response data to derive probabilistic thresholds for individual species may provide the most clear-cut robust approach to threshold derivation. The individual species may simply be the one that has been identified as the ‘most sensitive’ available indicator (in the case where 100% species protection is desired), or alternatively ‘keystone’ or ‘target’ species of high ecological relevance. The work of Fisher et al. (2018) represents an example of probabilistic thresholds for dredging related water quality parameters derived for ‘keystone’ taxa, based on an extensive field dataset, which shows how probabilistic thresholds can be used to derive different functionally useful thresholds within a conceptual decision science framework.

7.2 Multivariate methods

Moving from univariate data to multivariate data (data arrays involving multiple species and environmental variables) potentially provides enormous gains in sensitivity to ecosystem-level change across (and including contaminant) gradients. Multivariate statistical methods are rapidly evolving, and are highly popular in other related fields, such as community ecology. Their utility in the context of GV derivation has not been thoroughly evaluated, and there is considerable potential for their application in statistical ecotoxicology.

7.2.1 Similarity based community analysis

One recent approach to threshold derivation for field community data used a similarity-based, analysis approach, based on the differences between each level of contaminant concentration for each species. Specifically, the method compares similarity of (biological) compositional data within and between all possible pairs of derived classes of site contamination. While the method has been applied to field assemblage data (Kefford et al., 2010; Humphrey and Chandler, 2018), it can, in principle, be applied to laboratory toxicity data given that a suite of tested organisms can be considered a “community” (albeit limited species pool), with associated response to water quality. The laboratory-adapted approach lends itself to an analogous treatment-pairwise similarity approach from which a “community” CR model can be derived. The approach is appealing in:

- utilising fully *all* of the available CR data
- accounting for differences in responses amongst all possible pair-wise treatment comparisons using a well-founded (for biological data) distance measure (Bray-Curtis)
- negating the need for selecting single-species “no” or “low” effect concentrations (viz NOEC, NEC, EC10 etc) as a summary of the response of each species (see discussion in Section 4.1)
- by-passing the need for second-step SSDs which may be sensitive in outputs associated with (i) low species representation, and (ii) variable response data for relatively insensitive species.

An application of the method to toxicity data was demonstrated at the workshop with resulting GVs equivalent to those arising from the corresponding SSDs.

Analyses conducted to date are not sufficiently advanced to a point of a considered assessment of the validity and usefulness of similarity-derived GVs based on laboratory toxicity data. However, through comparative evaluations and simulations, the method is one amongst others that may assist in the process of identifying improved methods to determine reliably, GVs at low test species number.

Workshop attendees raised several issues they thought needed to be considered or addressed in assessing the potential of the similarity approach to laboratory data. Issues raised with post-workshop responses (from Chris Humphrey), included:

1. What does a change in similarity mean biologically?

Response: Similarity metrics are the basis of numerous multivariate methods that depict ecological distances between community samples. A change in metric value indicates a corresponding change or difference in community composition or structure between two samples and multivariate approaches are now standard and accepted as a valid method of depicting alterations in communities across contaminant gradients. These methods are not idiosyncratic or unusual in any way. As applied to laboratory toxicity data, such a change depicts a change in response to the contaminant at the ‘species pool’ level. Sensitive change detection per se amongst the tested organisms is at issue here.

2. Responses for adjacent concentrations appear to be more similar to one another as the concentrations increase. Further, none of the plotted data are truly independent of each other as the same data are used in sequential comparisons.

Response: The biological interpretation of the first observation needs to be assessed. Reducing the comparisons to just control to treatment, replicating this to the number of replicates used in a treatment would result in greater statistical independence.

3. SSDs never include toxicity estimates that are at, or are very close to, extrapolated 95 or 99% protection levels, whereas using the similarity method, interpolated PC95 or PC99 values appear to intersect actual (plotted (high) similarity values.

Response: The similarity ECO1 in the example provided at the workshop was higher than that derived from the corresponding SSD. However, this appears to arise from pooling similarities that include

those identified in 2 above – that have the effect of making the EC values less conservative. The refinement suggested in 2 would result in similar GVs (similarity vs SSDs).

4. *It was noted that concentrations don't always match up amongst test species, while the concentration to assign to a pairwise comparison needs some thought.*

Response: There will be ready solutions to these issues and they don't hinder further development of the concept.

7.2.2 TITAN

There are several statistical techniques that can identify tipping points or thresholds for species and communities to environmental variables. One such example is Threshold Indicator Taxa Analysis (TITAN) in R (Baker and King, 2010). This analysis examines how each taxon responds to an environmental gradient, in this case a contaminant. Using standardised z scores, TITAN can distinguish between those taxa for which their occurrence declines, and for those taxa which increase along the contaminant gradient. TITAN uses bootstrapping to estimate confidence limits around the change points, and the width of the confidence limits provides information on the tolerances of each taxon and community to the contaminant (Chariton et al., 2016). TITAN may be used to estimate a community-based guideline value, although it only indicates that there has been a change in community structure, not necessarily a change in the “health” of the community. It is also useful for selecting sensitive indicator species for bioassay development or biological monitoring, as additional lines of evidence in a WOE approach.

In a field-effects study conducted by Humphrey and Chandler (2018) amongst various methods to derive a threshold for magnesium (Mg) toxicity to macroinvertebrate communities in lentic waterbodies, TITAN was extremely conservative. The authors noted different peaks in taxa loss or gain, indicating different change points. Polymodality in change points was attributed mainly to missing (unmeasured) effects and concentration data for some intermediate Mg concentrations across the contaminant range. Thus at an assemblage level, taxa occurrence may be responding predominately to the frequency at which samples were collected and hence the different peaks may simply be artefacts of sampling intensity across the Mg concentration range. Thus, such bi- or polymodality in the peaks in taxa loss or gain may have resulted in the overly-conservative TITAN Mg change point for taxa disappearing, limiting the usefulness of the method. This phenomenon will likely be evident in any threshold method that does not relate biological response from exposed sites to contemporaneously measured responses from reference sites to standardise the response.

7.3 Recommendations

A thorough review of alternative methods to SSDs, in the context of evaluating their potential to support GV derivation and/or provide useful additional lines of evidence in WQ management, should be conducted.

8 SUMMARY

The Workshop resulted in the following recommendations:

SSD MODEL DISTRIBUTIONS AND MODEL FITS: Testing SSD fitting methods and the ssdtools Shiny app

1. There is a need to further investigate the methods used for AICc weighting and ensure that the most appropriate method for estimating weighted PC_x values is used. For example, should all the PC_x values be estimated and a weighted average calculated (as currently implemented in the ssdtools Shiny app), or should a mixture distribution be used to estimate PC_x directly (see above).
2. The models to be included in the candidate 'model set' need to be carefully considered. Models included must be 'plausible', meaning that they have the potential to provide realistic PC_x estimates and are commensurate with the nature and type of data. The Pareto model currently available in the ssdtools Shiny app would appear to be an example of a distribution that should be excluded as: (i) it is prone to stability issues in the estimation process, and (ii) it can yield non-sensical results (e.g. negative PC_x values). Furthermore, it is noted that the Pareto model is not currently included in the default set of candidate models in the ssdtools Shiny app. The models currently implemented in Burrlioz 2.0 should definitely be included, namely the log-logistic (currently one of the defaults in the ssdtools Shiny app) and Burr Type III. It was recognised that developing a set of criteria for model set selection would be beneficial.
3. It would be valuable to explore how the model averaging approach is affected by sample size, for a range of example data sets. AICc appropriately heavily penalises more complex models with very small samples. The current Australian and New Zealand GV derivation methods stipulate that a log-logistic distribution be used when there are toxicity data for <8 species, and a Burr Type III distribution be used when there are toxicity data for ≥8 species (Warne et al., 2018). It is worth investigating how the model averaging approach may align with this recommendation, and if small sample sizes do in fact favour simpler models (i.e. the log-logistic over the Burr Type III) and if so, at what sample size. Conversely, there may be some utility to including slightly more complex models (e.g. see Section 2.4 on Bimodality above) and testing that the model weights are robust with respect to 'overfitting' small samples. The heavy penalty against model complexity suggests that the model-averaging approach may be quite robust to the inclusion of more complex models, but this warrants thorough testing.
4. Some concern was expressed that having multiple very similar models in the model set may lead to an 'overweighting' of that functional form in the weighted averaging. Preliminary testing suggests that where similar models are included (for example the gamma and the

Gompertz) they effectively 'share' the model weight, suggesting that 'overweighting' is unlikely to be an issue, although this warrants further testing.

5. For the approach to be used for deriving GVs in Australia and New Zealand, clear guidance around the 'best practice' approach would need to be developed, and the current ssdtools Shiny app would need to be modified, possibly to include alternative 'tabs', e.g. a 'tab' for 'research mode' (perhaps similar to that already available) that is flexible and allows extensive exploration by the user, and a 'tab' that is clearly for use in official GV derivation, where the expected 'best practice' method is set by default, with minimum possible flexibility.

ACCOUNTING FOR UNCERTAINTY IN GUIDELINE VALUE DERIVATION: Capturing and reporting uncertainty in GV derivation

6. There is a range of existing methods for capturing uncertainty in underlying toxicity CR data that need to be thoroughly explored, including Monte Carlo simulation and Bayesian mixed modelling approaches. Possible methods need to be reviewed, and where appropriate compared, using real and/or simulated data.
7. Methods for allowing multiple equivalent endpoints to be included in the SSD need to be developed, so that their underlying uncertainty can also be taken into account.
8. While not explored further at the workshop, the utility of adopting approaches such as leave-one-out cross validation for capturing uncertainty around data-inclusion should be a target area for future investigation.
9. Regardless of the statistical methods used for capturing and propagating uncertainty in SSD estimation, more effort needs to go into developing user-friendly tools and interfaces, to increase accessibility to those without access to specialist statistical programming skills and to ensure the methods are accessible to a wide range of practicing ecotoxicologists.
10. More guidance on how to deal with the uncertainty and confidence limits is needed. Such guidance should include: (i) how confidence limits and uncertainty can be incorporated into management decisions through formal decision science (see Section 7.1), and (ii) wording around reporting to ensure confidence in the GVs and their subsequent regulatory use are maintained, despite the acknowledgement of (real) uncertainty.

INPUT TOXICITY DATA FOR SSDs

11. The following are required:
 - a. Straightforward robust guidelines on how best to estimate endpoints and accessible and easy to use tools to support accepted best methods for fitting CR datasets.

- b. Friendlier user interfaces for applying methods for calculating NEC and/or highly robust ECx values (perhaps developed through 'Shiny')
 - c. Guidance on designing ecotoxicological experiments that allow calculation of NEC and/or highly robust ECx values.
 - d. More discussion, guidance and/or 'rules' around the definition of 'acute' versus 'chronic' endpoints across a wider range of species and response variables.
12. Future emphases for species selection could potentially focus on development of test methods for typically under-represented taxa.

WEIGHTING OF INDIVIDUAL DATA POINTS / DATASETS

13. Although the current view is that there is insufficient basis upon which to defensibly adopt the use of weighting approaches for SSDs (Batley et al., 2018), more effort in this area might be warranted, including an assessment of any progress made over the past decade. Weighting for site- and application-specific GVs may be appropriate, and investigation around tools that allow this is warranted.

SMALL DATASETS

14. Explore a range of statistical approaches specifically in terms of their ability to cope with small sample sizes, including: model averaging, mixture modelling, uncertainty propagation, and cross validation.

ALTERNATIVES TO SSDs

15. A thorough review of alternative methods to SSD, in their context of evaluating their potential to support GV derivation and/or provide useful additional lines of evidence in WQ management, should be conducted.

9 REFERENCES

- ANZG (2018). *Australian and New Zealand Guidelines for Fresh and Marine Water Quality*. Australian and New Zealand Governments and Australian state and territory governments, Canberra, ACT, Australia. Available at www.waterquality.gov.au/anz-guidelines
- Baas J, Jager T, Kooijman B. 2010. *Understanding toxicity as processes in time*. *Science of the Total Environment* 408:3735–3739. doi:10.1016/j.scitotenv.2009.10.066. [accessed 2014 Dec 8]. <http://www.ncbi.nlm.nih.gov/pubmed/19969324>.
- Baker ME and King RS (2010). *A new method for detecting and interpreting biodiversity and ecological community thresholds*. *Methods in Ecology and Evolution* 1, 25-37.
- Barry S, Henderson B (2014). *Burrliz 2.0 CSIRO, Canberra, Australia*, Available from: <https://researchcsiro.au/software/burrliz/> Accessed December 24, 2014
- Batley GE, van Dam RA, Warne MStJ, Chapman JC, Fox DR, Hickey CW, Stauber JL Chapman JC, Fox DR, Hickey CW, Stauber JL (2018). *Technical rationale for changes to the method for deriving Australian and New Zealand water quality guideline values for toxicants— update of 2014 version*. Prepared for the revision of the Australian and New Zealand Guidelines for Fresh and Marine Water Quality. Australian and New Zealand Governments and Australian state and territory governments, Canberra, 43 pp.
- Burnham KP, Anderson DR (2002). *Model Selection and Multimodel Inference; A Practical Information-Theoretic Approach*. Springer, New York
- Burnham KP, Anderson DR (2004). *Multimodel inference: understanding AIC and BIC in model selection*. *Sociological Methods and Research* 33:261-304.
- Chariton AA, Pettigrove VJ, Baird DJ (2016). *Ecological assessment*. In Simpson SL and Batley GE (eds) *Sediment Quality Assessment: a Practical Guide*. 2nd edition, CSIRO Publishing, Victoria, pp 195-236.
- Dalgarno S (2018). *ssdtools: A shiny web app to analyse species sensitivity distributions* Prepared by Poisson Consulting for the Ministry of the Environment, British Columbia <https://poissonconsultingshinyappsio/ssdtools/>
- Dalgarno W, Cheng J, Allaire J, Xie Y, McPherson J. (2019). *shiny: Web Application Framework for R*. <https://CRAN.R-project.org/package=shiny>. Accessed 27/8/2019.
- ECCC (2018). *Biological Test Method: Test for Survival, Growth and Reproduction in Sediment and Water Using the Freshwater Amphipod Hyalella azteca*. Report RM/33, Third Edition - September 2017. Environment and Climate Change Canada. Gatineau, QC, Canada.
- Environment Canada (2017). *Biological Test Method: Test for Survival, Growth and Reproduction in Sediment and Water Using the Freshwater Amphipod Hyalella Azteca* Report RM/33 Third Edition - September 2017.

- Fisher R, Walshe T, Bessell-Browne P, Jones R, Trenkel V (2018). Accounting for environmental uncertainty in the management of dredging impacts using probabilistic dose-response relationships and thresholds. *Journal of Applied Ecology* 55:415-425.
- Forbes TL, Forbes VE (1993). A critique of the use of distribution-based extrapolation models in ecotoxicology. *Functional Ecology* 7:249–254.
- Forbes VE, Calow P (2002). Species sensitivity distributions: a critical appraisal. *Human and Ecological Risk Assessment* 8:473–492.
- Forbes VE, Calow P, Sibly RM (2008). The extrapolation problem and how population modeling can help. *Environmental Toxicology and Chemistry* 27:1987–1994. doi: 10.1177/0960327110391387.
- Forbes VE, Galic N, Schmolke A, Vavra J, Pastorok R, Thorbek P (2016). Assessing the risks of pesticides to threatened and endangered species using population modeling: a critical review and recommendations for future work. *Environmental Toxicology and Chemistry* 35:1904–1913. doi:10.1002/etc.3440.
- Fox DR (2010). A Bayesian approach for determining the no effect concentration and hazardous concentration in ecotoxicology. *Ecotoxicology and Environmental Safety* 73:123-131.
- Fox, DR (2015). Selection bias correction for species sensitivity distribution modeling and hazardous concentration estimation. *Environmental Toxicology and Chemistry* 34:2555-2563. <https://doi.org/10.1002/etc.3098>
- Fox DR, Landis WG (2016). Don't be fooled—A no-observed-effect concentration is no substitute for a poor concentration–response experiment. *Environmental Toxicology and Chemistry* 35:2141-2148.
- Fox DR (2018). *Contemporary Methods for Statistical Design and Analysis*. In: Blasco J, Chapman PM, Campana O, Hampel M (eds) *Marine Ecotoxicology*. Academic Press
- Gama-Flores JL, Sarma SSS, Nandini S (2005). Interaction among copper toxicity, temperature and salinity on the population dynamics of *Brachionus rotundiformis* (Rotifera). *Hydrobiologia*. 546:559–568. doi: 10.1007/s10750-005-4300-5.
- Gergs A, Gabsi F, Zenker A, Preuss TG (2016). Demographic toxicokinetic-toxicodynamic modeling of lethal effects. *Environmental Science and Technology* 50: 6017–6024. doi: 10.1021/acs.est.6b01113.
- Gottschalk F, Nowack B (2013). A probabilistic method for species sensitivity distributions taking into account the inherent uncertainty and variability of effects to estimate environmental risk. *Integrated Environmental Assessment and Management* 9:79-86.
- Hogan AC, Trenfield MA, Harford AJ, van Dam RA (2013). Toxicity of magnesium pulses to tropical freshwater species and the development of a duration-based water quality guideline. *Environmental Toxicology and Chemistry* 32:1969–1980
- Humphrey CL, Chandler L (2018). Use of field-effects information to derive a surface water guideline value for magnesium in Magela Creek, NT Australia. *Supervising Scientist Report 212*, Supervising Scientist, Darwin NT. <https://www.environment.gov.au/science/supervising-scientist/publications/ssr/use-field-effects-info>

- Jager T, Albert C, Preuss TG, Ashauer R (2011). General unified threshold model of survival - A toxicokinetic-toxicodynamic framework for ecotoxicology. *Environmental Science and Technology* 45:2529–2540. doi:10.1021/es103092a.
- Kefford BJ, Schäfer RB, Liess M, Goonan P, Metzeling L, Nugegoda D (2010). A similarity-index-based method to estimate chemical concentration limits protective for ecological communities. *Environmental Toxicology and Chemistry* 29:2123-2131.
- Kon Kam King G, Larras F, Charles S, Delignette-Muller ML (2015). Hierarchical modelling of species sensitivity distribution: Development and application to the case of diatoms exposed to several herbicides. *Ecotoxicology and Environmental Safety* 114:212–221. doi: 10.1016/j.ecoenv.2015.01.022.
- Mooney TJ, Pease CJ, Hogan AC, Trenfield M, Kleinhenz LS, Humphrey C, van Dam RA, Harford AJ (2019). Freshwater chronic ammonia toxicity: A tropical-to-temperate comparison. *Environmental Toxicology and Chemistry* 38:177-189.
- Newman MC, Ownby DR, Mézin LCA, Powell DC, Christensen TRL, Lerberg SB, Anderson BA (2000). Applying species-sensitivity distributions in ecological risk assessment: Assumptions of distribution type and sufficient numbers of species. *Environmental Toxicology and Chemistry* 19:508-515.
- OECD (2006). *Current approaches in the statistical analysis of ecotoxicity data: A guide to application*. OECD Environmental Health and Safety Publication, Series on Testing and Assessment, No. 54. Organisation for Economic Co-operation and Development, Environment Directorate, ENV/JM/MONO (2006) 18, Paris, France.
- OECD (2015). *The Larval Amphibian Growth and Development Assay (LAGDA)*. OECD Guidelines for the testing of Chemicals. Guideline 241, Organisation for Economic Cooperation and Development. Adopted 28 July, 2015, Paris, France.
- Peters A, Merrington G, Leverett D, Wilson I, Schlegel C, Garman E (2019). Comparison of the chronic toxicity of nickel to temperate and tropical freshwater species. *Environmental Toxicology and Chemistry* 38:1211-1220
- Pinheiro J, Bates D, DebRoy S, Sarkar D, R Development Core Team (2013). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-109.
- Pires AM, Branco JA, Picado A, Mendonça E (2002). Models for the estimation of a 'no effect concentration'. *Environmetrics* 13:15-27.
- Proctor A (2018). *Improvements in ecotoxicological analysis methods for the derivation of environmental quality guidelines: A case study using Antarctic toxicity data*. PhD Thesis. University of Tasmania
- Salice CJ, Miller TJ (2003). Population-level responses to long-term cadmium exposure in two strains of the freshwater gastropod *Biomphalaria glabrata*: Results from a life-table response experiment. *Environmental Toxicology and Chemistry* 22: 678–688.
- Schmolke A, Kapo KE, Rueda-Cediel P, Thorbek P, Brain R, Forbes V (2017). Developing population models: A systematic approach for pesticide risk assessment using herbaceous plants as an example. *Science of the Total Environment* 599–600: 1929–1938.

- Sinclair A, Tayler K, van Dam R, Hogan A (2014). *Site-specific water quality guidelines: 2. Development of a water quality regulation framework for pulse exposures of mine water discharges at a uranium mine in northern Australia. Environmental Science and Pollution Research* 21:131-140.
- Su Y-S, Yajima M (2015). *R2jags: Using R to Run 'JAGS'. R package version 0.5-6. <http://CRAN.R-project.org/package=R2jags>.*
- Swets JA, Dawes RM, Monahan J (2000). *Better decisions through science. Scientific American* 283:82-87.
- Thorley J, Schwarz C (2018). *ssdtools: Species Sensitivity Distributions R package version 003 <https://CRAN.R-project.org/package=ssdtools>*
- Thursby G, Sappington K, Etterson M (2018) *Coupling toxicokinetic-toxicodynamic (TK-TD) and population models for assessing aquatic ecological risks to time-varying pesticide exposures. Environmental Toxicology and Chemistry* 38: 2633–2644. doi: 10.1002/etc.4224.
- van Dam RA, Harford AJ, Warne MStJ (2012). *Time to get off the fence: the need for definitive international guidance on statistical analysis of ecotoxicity data. Integrated Environmental Assessment and Management* 8:242-245.
- Warne M, Batley GE, van Dam RA, Chapman JC, Fox DR, Hickey CW, Stauber JL (2018). *Revised Method for Deriving Australian and New Zealand Water Quality Guideline Values for Toxicants – update of 2015 version. Prepared for the revision of the Australian and New Zealand Guidelines for Fresh and Marine Water Quality. Australian and New Zealand Governments and Australian state and territory governments, Canberra, 48 pp.*
- Zajdlik BA, Dixon DG, Stephenson G (2009). *estimating water quality guidelines for environmental contaminants using multimodal species sensitivity distributions: a case study with atrazine. Human and Ecological Risk Assessment* 15:554-564.
- Zhao J, Zhang R (2017). *Species sensitivity distribution for pentachlorophenol to aquatic organisms based on interval ecotoxicological data. Ecotoxicology and Environmental Safety* 145: 193–199. doi: 10.1016/j.ecoenv.2017.07.029.

APPENDIX A: WORKSHOP ATTENDEES

Name	Affiliation	Primary expertise
Dr Rebecca Fisher	AIMS	Biostatistics/ecological modelling
Professor David Fox	Environmetrics Australia / University of Melbourne	Statistical science
Dr Abigael Proctor	AAD/UTAS	Biostatistics
Dr Quanxi Shao	CSIRO	Biostatistics
Patricia Menendez	AIMS	Biostatistics
Dr Rick van Dam	Independent consultant	Ecotoxicology
Dr Graeme Batley	CSIRO	Ecotoxicology/chemistry
Dr Michael Warne	UQ/Qld DES	Ecotoxicology
Dr Cath King	AAD	Ecotoxicology
Dr Andrew Harford	ERISS	Ecotoxicology
Dr Joost Van Dam	AIMS	Ecotoxicology
Dr Jenny Stauber	CSIRO	Ecotoxicology
Dr Andrew Negri	AIMS	Ecotoxicology
Dr Chris Humphrey	ERISS	Aquatic ecology

APPENDIX B: KEY ISSUES

Prior to the workshop, each attendee was asked to “please provide your list of 5 to 10 *key questions/problems/issues associated with current methods used for the derivation of water quality guidelines* (also referred to as standards or criteria)”. Responses to this request were collated into the 14 topic areas listed below.

SSD model distributions and model fits

Restricting the use of functions in Burrlioz might not be offering the best predictions, especially for small datasets, we should start including other model types. [Statistical]

Canadian ssdtools model-averaging approach

Bayesian model averaging approaches?

Is modality an issue? Can this be incorporated into the model averaging approach? [Statistical]

Should we be focussing more on getting a reliable fit at the bottom end of the SSD? [Statistical]

Assessment of the goodness of fit of the SSD curves. [Statistical]

At what stage should we say the predictions are too poor to be using SSD models and then we insist on other approaches [Guidance]

Uncertainty in GV derivation

Propagation of uncertainty around individual EC_x /NEC estimates within an SSD. [Statistical]

How to propagate uncertainty of endpoints converted between acute and chronic values? Or that are derived through means other than concentration response relationships? [Statistical]

How do we account for uncertainty associated with specifically included endpoints(species) and outliers, and should we? [Statistical]

How is uncertainty associated with the fitted model captured? Is this resolved through the use of model averaging? [Statistical]

How should a regulator use the 95% CI associated with published GVs and/or what national advice is provided on what the CIs mean or how they should be interpreted? [Guidance]

How we can statistically compare different PC95 values? [Statistical]

Without information from the 95% CI associated with published GVs, should another level of reliability be applied to derived GVs? [Guidance]

Input toxicity data for SSDs

Derivation and selection of toxicity estimates

Better experimental design: better replication, and dose response curves need better data for low response end. [Guidance]

Is EC10 a good/acceptable endpoint? Should we be striving for NEC or is a 10% effect tolerable?
What are the models/methods available? [Statistical] / [Guidance]

Where there are multiple experiments, should we be using an average ECx from replicate tests vs one ECx for pooled data from all replicate tests – is there criteria or a test to justify pooling data. There can be high variability for the same species among laboratories [Statistical] / [Guidance]

Does “% of control” in dose response relationships matter? Should we be normalising to controls? [Statistical] / [Guidance]

Temporal considerations

Integrating responses through time vs estimates at a particular exposure period/time [Statistical] / [Guidance]

We currently have no acute GVs. We need an approach that addresses the impact of acute toxicant concentrations and time of exposure (time averaged concentrations). [Data gap]

What about repeated exposure? No exposure studies consider recovery [Guidance]

Reliability of data and use of poor quality ecotoxicity data

There are fundamental questions about the data we are dealing with, their reliability and appropriateness for deriving protective concentrations. [Guidance]

Is it better to have more points and include low-quality data, or just stick with fewer high-quality data points? [Guidance]

While there is a well-established system for scoring the quality of underlying toxicity data, these scores are not currently used to weight the data’s contribution to the SSD. Should this be done? If so, is there already methods for doing this (within an SSD context, or otherwise)? [Statistical]

Weighting of individual data points/datasets

Inclusion of weighting of individual data points in an SSD (especially with small data sets) based on

1. certainty of value,
2. ecological importance,
3. single vs community response,
4. rapid vs conventional testing

This is particularly relevant for small data sets. [Statistical]

Taxonomic/regional ecological relevancy of the datasets

How does the species selection impact guideline values? [Statistical] / [Guidance]

Fundamental objectives of “the derivation of water quality guidelines”.

Sensitive species or ‘keystone’ species? [Guidance]

Tropical v temperate v Antarctic? [Statistical]/ [Guidance]

Different water bodies/regions [Statistical]/ [Guidance]

Small datasets

What are the best practice approaches for deriving GVs when sample size is small? Should approaches alternative to SSDs be considered? What is a minimum? [Guidance]

When sample sizes are small individual values can have too large a relative impact. Is there a way to deal with this? [Statistical]

For ecosystems that require high levels of protection (e.g. PC99), are these values on the extreme tail of the distribution even valid for small sample sizes? Very few data are being used to estimate guideline values on the extreme tail of the SSD. Are there better alternative methods? [Statistical]

At what stage should we say the predictions are too poor to be using a SSD models and then we insist on other approaches. [Statistical]/ [Guidance]

Alternatives to SSDs

Use of individual dose-response values for keystone species

Field or mesocosm community assessments for GV derivation

Methods employing all of the available dose-response data from all species tested, negating the need for SSDs, provide values less prone to variation associated with SSDs based on small test species number or outliers/extremes? [Statistical]

Probabilistic thresholds based on aggregated probability distributions of endpoint values [Statistical]

How to we integrate/incorporate other information (lab/field/expert opinion)

Are there ways to quantitatively integrate other types of data into SSDs? [Statistical]

- Field data
- Non-standard data (community metrics, model outputs)
- Expert opinion

Should non-standard data be weighted differently to standard data? [Statistical] / [Guidance]

Exposure mechanisms/characteristics

How do we deal with complex mixtures and multiple stressors? [Guidance]

How best to incorporate other factors affecting toxicity and speciation? [Guidance]

How do we best include bioavailability into guideline derivations and SSDs?

How do we select the best model (BLM or MLRs etc) for normalisation of metals ecotox data?

Do we need to normalise at all for water chemistry? [Guidance]

How to better deal with species/ecosystem interactions? [Guidance]

How best to incorporate different exposure pathways (especially dietary vs absorption which is normally what is tested for in aqueous toxicity tests but may not be the primary route of exposure).

What about bioavailability and importance of the food chain [Guidance]

Implementation/application of guidelines

Better education in the implementation of GVs [Guidance] Minimum standards syndrome.

The “gold standard” syndrome. DGVs are often implemented as pass/fail gold standards.

Overall, inclusion of SSD findings should be considered as just one line of evidence for toxicity.

Statistical practice in ecotoxicology

Form guidelines in the use of statistics with a framework to follow. [Guidance]

Deciding which methods suit which situations – i.e. not being limited to just one method, but several that are suited to specific types of toxicants or datasets. [Guidance]

Uptake - While statistical methodology does exist to cope with many of the inadequacies of current practice, eco-toxicologists aren't statisticians. We need user friendly and robust approaches that are accessible. We need documentation that is clearly justified and understandable, at least conceptually, by non-statisticians. [Statistical] / [Guidance]