Joe Thorley ORCID iD: 0000-0002-7683-4592

20-00450

Critical Review

Recent developments in SSD Modeling

D.R. Fox[* 1,2], R.A. van Dam[3], R. Fisher[4], G.E. Batley[5], A.R. Tillmanns[6], J. Thorley[7], C.J. Schwarz[8], D.J. Spry[9] and K. McTavish[9]

[1] Environmetrics Australia Pty Ltd, Beaumaris, Vic 3193, Australia

[2] University of Melbourne, Parkville 3010, Australia

[3] WQadvice, Adelaide, SA, Australia

[4] Australian Institute of Marine Science, Crawley, WA, Australia & the UWA Oceans Institute and School of Plant Biology, University of Western Australia, Crawley, WA, Australia

[5] CSIRO Land and Water, Lucas Heights, NSW, Australia

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/etc.4925.

[6] BC Ministry of Environment and Climate Change Strategy, Victoria, BC, Canada

[7] Poisson Consulting Ltd., Nelson, BC, Canada

[8] StatMathComp Consulting, Vancouver, BC, Canada

[9] Environment and Climate Change Canada, Gatineau, QC, Canada

(Submitted 22 June 2020, Returned for Revisions 13 July 2020, Accepted 30 October 2020)

Abstract

The species sensitivity distribution (SSD) is a statistical approach that is used to estimate either the concentration of a chemical that is hazardous to no more than x% of all species (the HCx) or the proportion of species potentially affected by a given concentration of a chemical. Despite a significant body of published research and critical reviews over the past 20 years aimed at improving the methodology, the fundamentals remain unchanged. While there have been some recent suggestions for improvements to SSD methods in the literature, in general, few of these suggestions have been formally adopted. Further, critics of the approach can rightly point to the fact that differences in technical implementation can lead to marked differences in results, thereby undermining confidence in SSD approaches. Despite the limitations, SSDs remain a practical tool and, until a demonstrably better inferential framework is available, developments and enhancements to conventional SSD practice will and should continue. We therefore believe the time has come for the scientific community to decide how it wants SSD methods to evolve. The current paper summarises the current status of, and elaborates on

several recent developments for, SSD methods, specifically, model averaging, multimodality and software development. The paper also considers future directions with respect to the use of SSDs, with the ultimate aim of helping to facilitate greater international collaboration and, potentially, greater harmonization of SSD methods.

Keywords: Species sensitivity distrbution; statistical inference; hazardous concentration; computer software.

INTRODUCTION

The species sensitivity distribution (SSD) is a statistical approach that is used to estimate either the concentration of a chemical that is hazardous to no more than x% of all species (the HCx) or the proportion of species potentially affected by a given concentration of a chemical. Following its introduction in the 1980s (Stephan et al. 1985; Kooijman 1987; van Straalen and Denneman 1989), the species sensitivity distribution (SSD) remains the most widely used method for deriving water quality benchmarks (guidelines, criteria or standards, depending on the jurisdiction) to characterize effects of chemical contaminants for water quality and/or ecological risk assessment purposes. The SSD has proven to be a useful, practical and intuitive tool (see Belanger et al. 2017, 2019), albeit not without numerous limitations (e.g., OECD 1992; Forbes and Forbes 1993; Smith and Cairns 1993; Warne 1998; Newman et al. 2000; Forbes and Calow 2002; Wheeler et al. 2002a,b; Zajdlik 2006; Hickey and Craig 2012; ECETOC 2014), including the implausibility of the many assumptions underpinning SSDs and concerns arising from inconsistent statistical results. Despite a significant body of published research and numerous intensive reviews (e.g., OECD 1992; Posthuma et al. 2002; ECETOC 2014; Fisher et al. 2019) over the past 20 years aimed at improving SSD methods, the fundamental SSD

approach employed by jurisdictions around the world has remained similar. However, variations do exist in some of the technical details of the methods and associated software tools that have been developed and employed, which can lead to marked differences in results and undermine confidence in SSD approaches.

Despite the limitations, SSDs remain a practical tool and, until a demonstrably better inferential framework is available, developments and enhancements to conventional SSD practice will and should continue. Indeed, numerous studies have attempted to address many of the limitations, including issues of sample size, species representativeness and selection, test endpoints, ecological relevance, phylogenetic relatedness and routes of exposure (e.g., de Zwart and Posthuma 2005; Dyer et al. 2006; Fox 2010; Wang et al. 2015; Warne et al. 2018; Belanger and Carr 2019; Carr and Belanger 2019; Moore et al 2019; Schwarz and Tillmanns 2019). While there has been some recent adoption of improvements to formal SSD methods (i.e., methods typically approved and recommended for use by national, provincial and state regulatory bodies) (e.g., Warne et al. 2018; British Columbia Ministry of Environment and Climate Change Strategy 2019), in general, few of the outcomes of SSD studies from the past 20 years have been formally adopted. Moreover, where refinements to formal SSD methods have been made, they have typically been done at a national or regional scale and over different timeframes, and in the absence of any globally agreed consensus or vision. We believe the time has come to stand back and assess what has been done to date and how, as a scientific community, we want SSD methods to evolve.

The current paper summarizes the current status of, and elaborates on some specific, recent developments for, SSD methods, specifically, model averaging (where the HCx is

estimated using a weighted-average of a number of individual SSDs), multimodality and software development. The paper also considers future directions of the use of SSDs, with the ultimate aim of helping to facilitate greater international collaboration and, potentially, greater harmonization of SSD methods.

CURRENT STATUS

SSD methodologies

This section provides a brief summary of the history and progress of formal SSD methods in key jurisdictions.

The history of the application of SSDs in North America has been well documented by Suter (2002) and by Stephan (2002). The current method in the U.S. for deriving water quality benchmarks (WQBs) (Stephan et al. 1985) has been in place for 35 years. To derive the HC5, a log-triangular distribution is applied to the four genus-level toxicity values whose cumulative probabilities are closest to the 0.05 probability point, which, except for very large data sets, will always correspond to the four most sensitive genera. Long-awaited revisions to the approach of the USEPA are embodied in the recently released SSD Toolbox software (Center for Computational Toxicology and Exposure 2020). However, SSD Toolbox is not an update to the USEPA's long-standing WQBs derivation methodology (Stephan et al. 1985), but instead has been developed to allow users to use statistical methods and approaches that reflect their risk assessment objectives (M. Etterson, USEPA, pers. comm). SSD Toolbox, also incorporates model averaging similar to the approach developed in Canada (see below).

SSD-based approaches have been used by various European countries since the 1980s for both WQB derivation and risk assessment purposes. A harmonized approach for deriving WQBs, which included the use of SSDs, was adopted across the European Union (EU) in 2005 (Lepper 2005) and updated in 2011 (European Commission 2011). The approach permits the use of different parametric distributions (e.g., log-normal, log-logistic, Burr Type III) for the SSD, but requires thorough justification if the choice of distribution is not the log-normal or log-logistic. The use of the ETX-computer program (Van Vlaardingen et al., 2004) is recommended as being appropriate for calculating HCx values, although it is not prescribed. Another key SSD software tool, developed in France, is MOSAIC (Kon Kam King et al. 2014).

In 2000, Australia and New Zealand (ANZECC/ARMCANZ 2000) adopted an SSD-based method for deriving WQBs, following a critical review of multiple WQB derivation methods (Warne 1998). A distinct feature of the method was the use of a three-parameter Burr distribution to model the empirical SSD, which was implemented in the Burrlioz software tool (Campbell et al. 2000). This represented a generalisation of the methods previously employed by Aldenberg and Slob (1993) since the log-logistic distribution was shown to be a specific case of the Burr family (Tadikamalla 1980). Recent revision of the derivation method recognized that using the three-parameter Burr distributions for small sample sizes (<8 species) created additional uncertainty by estimating more parameters than could be justified, essentially over-fitting the data (Batley et al. 2018). Consequently, the method, and updated software (Burrlioz 2.0), now uses a two-parameter log-logistic distribution for these small data sets, while the Burr

type III distribution is used for data sets of 8 species or more (Batley et al. 2018; ANZG 2018).

In Canada, the transition from a deterministic approach to the preferential use of SSDs occurred in 2007 (CCME 2007). Reviews of available statistical models by Zladjik (2005, 2006) recommended the choice of a single statistical distribution from a suite of at least six distributions (i.e., Burr Type III, Gumbel, Logistic, Log-normal, Normal, and Weibull), with goodness-of-fit analysis used to determine the most appropriate model. This was implemented in SSD Master (CCME 2013), an Excel macro, which uses ordinary least squares to fit a SSD to the empirical cumulative distribution function (*cdf*). This contrasts with most other methods that use maximum likelihood estimation (MLE). More recently, the BC Ministry of Environment and Climate Change Strategy developed a model-averaging approach using the R package ssdtools (Thorley and Schwarz 2018), and this has been used at the national level (CCME 2019, 2020). A web-based app, shinyssdtools, has also been developed to provide a Graphical User Interface (GUI) for the ssdtools R package (Dalgarno 2018). Hereafter, we use the term (shiny)ssdtools to refer to both ssdtools and shinyssdtools.

Currently available SSD software tools

Currently, there are at least nine software tools for fitting SSDs using a variety of methods (Table 1). We consider Maximum-Likelihood to be the most suitable method from a regulatory perspective, because it is generally less biased than moment matching, does not require the specification of prior distributions (unlike Bayesian methods) and lends itself to model averaging (unlike Least Squares).

The tools, which are free to use, all estimate the HC5, and most will estimate an HCx for any user-supplied value of *x*, together with confidence intervals. The most common distributions are the log-logistic and log-normal, which are each implemented in six of the nine tools. All the distributions are two parameter distributions (the log-triangular is symmetric) except for the Burr Type III and the log-t distributions. Four of the software tools in Table 1 (hSSD, MOSAIC, SSD Toolbox and (shiny)ssdtools) handle censored data which is an important facility when dealing with small data sets that contain observations expressed as "<" or ">" values (see Kon Kam King et al. 2014; Aldenberg 2015).

Only the SSD Toolbox (Etterson 2020), which has recently been released by the US EPA, and (shiny)ssdtools, which was developed for the British Columbia Ministry of Environment and Climate Change Strategy (Thorley and Schwarz 2018; Dalgarno 2018), implement model averaging. It is important to be aware that (shiny)ssdtools consists of ssdtools - a stand-alone R package (Thorley and Schwarz 2018) - and shinyssdtools (Dalgarno 2018) a second R package which provides a bilingual (English and French) GUI to ssdtools. The advantage of this separation is discussed below. As only SSD Toolbox and (shiny)ssdtools fit six of the 10 distributions (see Table 1) using maximum likelihood, run on all three major platforms (see Table 1), and have GUIs, we consider them to be the most useful of the nine software tools. Consequently, they are the focus for the remainder of this section.

SSD Toolbox is written in the commercial MATLAB® language and provided as a pre-compiled binary that can be run by locally installing the free MATLAB® Runtime libraries. ssdtools and shinyssdtools are both written in the open source R language (R

Core Team 2020). The source code for both has been released under the open source Apache-2.0 Licence (https://github.com/bcgov/ssdtools and https://github.com/bcgov/shinyssdtools), which allows users to modify and/or distribute the code under the same licence. We consider open source software to be preferable to compiled code because it allows code validation and facilitates collaboration and replication (Munafò et al. 2017; Mancini et al. 2019). SSD Toolbox allows distributions to be fitted using Bayesian methods and can statistically account for multiple datapoints for each species using hierarchical models. Neither of these features is currently implemented in (shiny)ssdtools. However, by separating the scripting and GUI components into ssdtools and shinyssdtools, respectively, developers can readily extend (shiny)ssdtools functionality or incorporate it into their own software. Shinyssdtools also provides an R script allowing the user to replicate the analysis they performed through the GUI. Finally, a web-based version of shinyssdtools which does not require the user to install R and runs on any browser is available at https://bcgov-env.shinyapps.io/ssdtools.

While we are not advocating adoption of a single standard approach or software tool, we think there is a need for closer jurisdictional collaboration, greater harmonisation of methods, and development of at least some benchmark data sets and reference results. The last of these is particularly pressing given the frequency with which we have observed noticeably different HCx values for the same data set from the different tools in Table 1. Although outside the scope of the current review, a comprehensive review of features together with detailed performance comparisons is currently being prepared for a follow-up paper. Some differences between the outputs of different tools are to be expected if different estimation strategies are employed (for example, maximum

likelihood versus moment matching or single SSD versus a model-averaged SSD) but all things being equal, all tools should give the same point estimates to within some nominally small tolerance (e.g., 1-2%). Certainly, differences of a factor of 2 or more are indicative of flawed coding and/or numerical instabilities and convergence issues.

The use of 'reference data sets' is not a new idea; they were commonly used in the early days of statistical computing to allow both software developers and end-users to assess the adequacy of numerical routines underpinning routine analyses such as ANOVA, regression, and correlation. Even today, the National Institute of Standards and Technology still maintains a number of statistical reference data sets at https://itl.nist.gov/div898/strd/index.html, including the famous Longley data set (Longley 1967).

TECHNICAL CHALLENGES

There have been several improvements to the SSD methodology over the last 20 or so years, most of which have been driven by advances in computing technology. For example, the early preferential use of the log-logistic distribution as a candidate SSD was not because it is intrinsically better than alternative distributions such as the log-normal, gamma or 3-parameter Burr distribution, for example, but because "it has some nice mathematical features that make certain calculations relatively easy" (Aldenberg and Slob 1993). Software tools like SSD Master (Intrinsik 2013) utilised the simplicity and computational power of Excel to fit a wider range of theoretical probability distributions. However, the lack of more sophisticated algorithms in Excel meant that this was done using statistically inferior methods to the generally preferred maximum likelihood estimation procedure. Most contemporary software tools for SSD modelling utilise a

combination of maximum likelihood estimation of the HCx and resampling methods such as the bootstrap (Efron and Tibshirani 1986) to obtain confidence intervals. Alternative statistical paradigms, such as Bayesian methods are now viable alternatives for ecotoxicology (Fox 2010; Zhang et al. 2012) because of the ready availability of free software tools such as JAGS (Plummer 2003) and STAN (Gelman 2015) coupled with the computational power of modern desktop computers. The use of non-parametric or 'distribution-free' statistical methods has been suggested as a means of overcoming the drawbacks associated with fitting and using SSDs (Carr and Belanger 2019; Van Der Hoeven 2001), although as noted by Van Der Hoeven (2001), such methods are unlikely to be useful for n<19. For samples of size 20 or more, parametric modelling of the SSD as discussed in this paper, provides a richer statistical framework than non-parametric counterparts.

Against this backdrop of continual refinement and improvement, SSD modelling continues to be hampered by some persistent and seemingly intractable problems. Deficiencies in the theory and application of SSDs have been comprehensively documented in the literature and while it is not our intention to revisit those here, the critical issue of identification of the functional form of the SSD represents an ongoing challenge for ecotoxicology.

Identification of the functional form of the SSD

Many authors have noted that there is no guiding theory in ecotoxicology to justify any particular distributional form for the SSD other than its domain be restricted to the positive real line (Fox 2016; Newman et al. 2000; Zajdlik 2005; Chapman et al. 2007).

Indeed, Chapman et al. (2007) described the identification of a suitable probability model as one of the most important and difficult choices in the use of SSDs. Compounding this lack of clarity about the functional form of the SSD is the omnipresent, and equally vexatious issue of small sample size. As noted by Chapman et al. (2007) and Fox (2016), small samples result in low-powered goodness-of-fit tests meaning *any* plausible candidate model is unlikely to be rejected by these procedures. For example, consider the small toxicity data set: {13.26, 8.27, 21.22, 16.23, 17.02, 3.28}. Plots and Anderson Darling goodness-of-fit test statistics suggest that the lognormal, Weibull, log-logistic, and gamma distributions are all plausible SSDs for these data (Figure 1).

The fixation on distributional form and fit is somewhat unique to SSD modelling because it defines the behavior of the model(s) in the left-tail region of the distribution. To illustrate why this is important, consider the triangular and log-normal distributions in Figure 2a. The distributions have the same means (7.5) and same variances (19.5) and are both positively skewed. However, in the region of interest to ecotoxicologists, these two distributions are very different (Figure 2b). Unlike the log-normal distribution, the triangular distribution has an abrupt cut-off resulting in very different probability and quantile determinations. Thus, for the distributions in Figure 2, we would conclude that the fraction of affected species at a concentration of 1.5 units is either 3% using the triangular distribution or 0.4% using the log-normal distribution.

Clearly, modelling the left tail in a manner that most closely resembles the underlying but unknown distribution is of critical importance in ecotoxicology, yet in practice this is precisely the region of greatest uncertainty.

Multimodality

In our experience, multimodality (and in particular, bimodality) of the empirical SSD is not uncommon. This arises because the toxicity data underpinning the empirical SSD are *not* from a single, common probability model as is conventionally assumed. The use of toxicity data that relate to different taxonomic groups, endpoints, test durations, modes of action, or sensitivities will often result in multimodal SSDs. At the very least, a somewhat arbitrary dichotomy is usually identified based on test organisms being 'more' or 'less' sensitive to the toxicant under consideration.

As an obvious example, the toxicity of a herbicide to plants and animals will, by design, often be markedly different. In such cases, the empirical SSD will exhibit *bimodality* (that is, the distribution has two modal values). We acknowledge that, in a regulatory risk assessment context, the usual advice is to fit different SSDs to different segments of species at risk (e.g. aquatic plants, terrestrial plants, aquatic invertebrates and fish). Such advice is sound when there is a clear regulatory interest in the different species groupings. However, as previously noted, this may be impractical if insufficient data are available in any or all subgroups to meaningfully fit an SSD. Further, as we do not know the exact mode of action of many substances, bimodality cannot be ruled out for any dataset, and it may also be unclear which endpoint observations belong to which group where there is overlap in the multiple distributions. Finally, in the derivation of WQBs, it is a common and accepted practice internationally to derive concentrations of chemicals that will protect (or not affect) aquatic ecosystems as a whole, not just a specific sub-group of the ecosystem, thus methods for estimating percentage species protection values that can accommodate multimodality of the underlying distribution are needed.

While theoretical bimodal univariate probability distributions do exist (for example Equation 1 and Figure 3), these are relatively uncommon and lack the flexibility to be useful candidate SSDs.

$$f(x;\mu,\sigma) = \frac{\left[1+\left(\frac{x-\mu}{\sigma}\right)^2\right]\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}{7\sqrt{\frac{\pi}{2}}} \quad (1)$$

Although there is little guidance on effective strategies for SSD modelling with multimodal data sets, the recently revised Australian and New Zealand method for deriving guideline values (GVs) for toxicants (Warne et al. 2018) suggest taking a weight-of-evidence (WoE) approach based on a chemical's mode of action, indications of bimodality or multimodality, and the presence of taxa-specific sensitivity. Further, Warne et al. (2018) suggest indications of bimodality should be based on both a visual inspection of the empirical SSD coupled with the computation of the bimodality coefficient (BC) (Freeman and Dale 2013; Pfister et al. 2013) given by Equation 2.

$$BC = \frac{\gamma^2+1}{\kappa+\frac{3(n-1)^2}{(n-2)(n-3)}} \quad (2)$$

where γ is the skewness; κ is the excess kurtosis; and n is the sample size.

Although there is no formal test of significance associated with the bimodality coefficient, a rule-of-thumb is that a value exceeding 0.555 (the value for a uniform distribution) is consistent with an underlying bimodal SSD (Freeman and Dale 2013; Pfister et al. 2013).

When the WoE assessment indicates the presence of bimodality that is known or thought to be due to a specific mode of action, the recommendation is that "the data set should be split and only the data belonging to the most sensitive group of species should be used to derive the GV" (Warne et al. 2018). This is also consistent with the advice given by Stephan et al. (1985). If the bimodality cannot be linked to a specific mode of action the recommendation is to use professional judgement, although what that might entail is not specified.

Splitting a small toxicity data set into even smaller subsets based on either known or assumed toxicity groupings is only feasible when the number of toxicity values in each of the subsets satisfies recommended minimum sample size requirements for SSD modelling. Failing this, the researcher is presumably left with the choice of either fitting a single SSD to the complete (bimodal) data set or abandoning the SSD modelling exercise altogether.

The first option leads us to reflect on the desirability of fitting a theoretical distribution using criteria that aim to minimize the disparity between the empirical and fitted distributions over the *entire range of toxicity values*. These so-called 'goodness-of-fit' measures are sensible and work well for most applications of statistical distribution-fitting. However, as noted above, our ultimate use of the SSD is restricted to a very narrow (lower left) portion of the domain. This has led to suggestions such as fitting the SSD using a method that somehow gives more weight to the data in the left tail of the SSD or, alternatively, fitting a mixture of different distributions (see *Dealing with Multimodality* below).

RECENT DEVELOPMENTS

Several meetings have been held to assist in the identification and resolution of some of the more substantive issues in SSD modelling. ECETOC (2014) represents the most recent global scale effort to review how SSD methodologies could be further developed. It was attended by 41 experts from academia, government, and industry from 13 countries, and provided a useful snapshot of some of the more recent developments in SSD methodologies, including inter-species correlation estimation, field-based SSDs and Bayesian approaches. Belanger et al. (2017) provided a useful summary of the key issues discussed at, and research needs arising from, the workshop. However, ECETOC (2014) was a one-off workshop and, although there was interest in a coordinated ongoing work program, there has been no subsequent coordinated forum to continue discussions on the advancement of SSD methodologies. Recently, and divergent from the typical thinking around the construction of SSDs and derivation of WQBs, Posthuma et al. (2019) suggested the need to relax the strict criteria typically prescribed for data selection and SSD construction in order to derive WQBs and assess risk for many more chemicals than is currently the case. Applying such an approach, they constructed SSDs for over 12,000 chemicals, adding to the debate on whether it is better to have more reliable SSDs and WQBs for fewer chemicals or potentially less reliable SSDs and WQBs for many more chemicals.

In March 2019, a group of 14 Australian scientists met at the Australian Institute of Marine Science in Townsville, Queensland for 3 days to discuss options for improving SSD methodologies for deriving water quality benchmarks (Fisher et al. 2019). Subsequently, in November 2019, key Australian and Canadian researchers met in

Victoria, BC (Canada) to identify commonalities in SSD research with a view to harmonizing strategies and approaches to SSD methodology development. Both meetings paid particular attention to model averaging and statistical mixture modelling (discussed below). These are promising new developments, the first of which has been championed by Schwarz and Tillmanns (2019) while the second was heavily promoted by Fox (Fisher et al. 2019).

Our intention here is not to chronicle all recent advances, but rather to highlight a smaller number of newer developments and opportunities that address some persistent and problematic issues with fitting and using SSDs. We consider: (i) model averaging as a means of alleviating problems with choosing a single probability model; (ii) statistical mixture modelling to overcome issues associated with bi- and multi-modality of the empirical SSD; and (iii) weighting of the lower tail of the SSD to better reflect our interest in this portion of the SSD. The material in the following section draws heavily from the Townsville workshop (Fisher et al. 2019) and the work undertaken by the BC Ministry of Environment and Climate Change Strategy (Schwarz and Tillmanns 2019).

Model averaging

The absence of biological theory coupled with equivocal statistical guidance has prompted researchers to consider alternative approaches to SSD model identification. One such option is *model averaging*, which potentially provides a more objective (and robust) way of handling the uncertainty associated with the identification of the appropriate distributional form for the SSD.

Model averaging is an alternative strategy to picking a single 'best' distribution and has recently been adopted by the BC Ministry of Environment and Climate Change Strategy (2019) through the use of ssdtools and is also an option within the USEPA's recent SSD Toolbox software (Center for Computational Toxicology and Exposure, 2020). Schwarz and Tillmans (2019) used data sets extracted from the CCME Guidelines for the Protection of Aquatic Life for boron (CCME 2009) and silver (CCME 2015) to assess and compare the results from model averaged and single SSDs. Among other things, they concluded that model averaging can reduce the uncertainty associated with fitting distributions to small data sets as well as providing some immunity to perturbations in HCx values due to the influence of a single sensitive data point.

The idea of model averaging is straightforward and is analogous to any averaging process that aims to 'iron out the bumps'. For example, consider the familiar problem of estimating the true mean of a variable, *X*. Standard statistical theory tells us that if $\{X_1,...,X_n\}$ are independently, identically distributed random variables with mean μ and variance σ^2 then the sample mean, $\bar{X}$ has the same mean μ but variance $\frac{\sigma^2}{n}$. So while both the sample mean and an *individual* observation are unbiased estimators for the true mean μ, $\bar{X}$ is better in the sense that its variance is $\left(\frac{1}{n}\right)^{th}$ that of *X*. In other words, $\bar{X}$ is more *precise*.

Just as a sample average provides a statistically better estimate of the true mean than any individual observation, we expect that the average of several estimates of an HCx from a set of plausible SSDs to do better than any individual estimate from a single SSD. In this

case, 'better' means the variance of the error associated with a model-averaged estimate is less than that from any single model. While this generally turns out to be the case, it has been suggested that model averaging is only likely to be useful when the error of contributing model predictions is dominated by variance, and if the covariance between models is low. (Dormann et al. 2018).

Another potentially problematic issue with model averaging is the selection of candidate probability models. As noted by Burnham and Anderson (2002), the construction of the candidate model set involves an element of subjectivity and that "one must recognize a certain balance between keeping the set small and focused on plausible hypotheses, while making it big enough to guard against omitting a very good a priori model". Similarly, Wheeler and Bailer (2009) pointed out that the efficacy of model averaging when applied to dose-response modelling is very much dependent on the model space (i.e. the set of candidate models).

As alluded to at the beginning of this section, the central issues in model averaging are bias and precision (reciprocal of variance). While a full mathematical treatment of model averaging is outside the scope of this paper, a brief review will give context to the remainder of the discussion.

We commence by letting $\hat{T}_m$ denote an estimator for some parameter θ (e.g., the HCx from model *m*. An assessment of the adequacy of $\hat{T}_m$ as an estimator of θ is provided by the mean square error (MSE) defined as follows:

$$MSE\left(\hat{T}_m\right) = E\left[\left(\hat{T}_m - \theta\right)^2\right] = \left\{\text{bias}\left(\hat{T}_m\right)\right\}^2 + Var\left[\hat{T}_m\right] \quad (3)$$

where $E(\cdot)$ in Equation 3 denotes mathematical expectation. The simplest model-averaged estimate of θ (denoted $\hat{\theta}$) from *k* models is the arithmetic mean:

$$\hat{\theta} = \frac{1}{k}\sum_{m=1}^{k}\hat{T}_m \quad (4)$$

Equation 4 is a specific case of the more general weighted average in which each model-averaged estimate is assigned the same weight of $w_m = \frac{1}{k}$ in Equation 5.

$$\hat{\theta} = \sum_{m=1}^{k} w_m \hat{T}_m; \quad \text{where } 0 \le w_m \le 1 \text{ and } \sum_{m=1}^{k} w_m = 1 \quad (5)$$

In the context of SSD model averaging, the assignment of equal weights would rarely make sense since it would be inconsistent with both subjective assessment and statistical measures of goodness-of-fit; namely, not all SSD models perform equally well in describing a given toxicity data set. So, the issue becomes one of selecting an 'optimal' set of w_m values in Equation 5. But, as noted by Dormann et al. (2018), estimation of this 'optimal' set of weights is itself subject to uncertainty – we don't know the *true* values of this optimal set that yield the smallest MSE. In other words, the *estimated* 'optimal' weights will be suboptimal, meaning the use of Equation 5 with weights estimated from the data may result in an estimate that is no better than one obtained

using arbitrary weights, e.g., equal weights (Dormann et al. 2018).

Although there are various strategies for estimating the weights in Equation 5, perhaps the most common are those based on information-theoretic concepts such as the Kullback-Leibler divergence which, loosely speaking, is a measure of the 'distance' between a given model and a reference model (Kullback 1959). We will restrict our attention to just one of these measures - Akaike's Information Criterion or *AIC* (Akaike 1973) given by Equation 6.

$$AIC = 2p - 2\ln(\hat{L}) \qquad (6)$$

where p is the number of parameters in the model and $\hat{L}$ is the maximum value of the likelihood function for the model. When n, the number of samples is small and the models have different numbers of parameters, then the following corrected version (AIC_c) of the *AIC* is preferred:

$$AIC_c = AIC + \frac{2p(p+1)}{n-p-1} \qquad (7)$$

Note that the value of *AICc* converges to the value of *AIC* for an infinitely large sample size. In the context of ecotoxicology, sample sizes are almost invariably considered 'small' (n/p<40, Burnham and Anderson 2002), thus *AICc* should generally be used in the SSD context. An exception is when the data are arbitrarily censored (i.e., where a cut-off is recorded rather than an actual numerical value) so as to give more weight to the left tail as described in the section *Left Tail weighting of the SSD* below. In such cases, *n* is no

longer defined and AIC should only be used if the models have the same number of parameters. Hereafter, we refer to either version as simply an *AIC*.

By itself, the *AIC* is not particularly useful. Its primary role is to assess and rank a series of candidate models. This is done by forming the *AIC differences*:

$$\delta_m = AIC_m - AIC_{\min} \quad (8)$$

where $AIC_{\min}$ is the smallest *AIC* in the set of *k* models. In broad terms, the empirical support is: high for models with $\delta_m \le 2$; substantially less for models with $4 \le \delta_m \le 7$; and virtually nil for models with $\delta_m > 10$ (Burnham and Anderson 2002).

The weight to be assigned to the estimate from model *m* is then computed using Equation 9.

$$w_m = \frac{e^{-\frac{1}{2}\delta_m}}{\sum_{i=1}^{k} e^{-\frac{1}{2}\delta_i}} \quad (9)$$

If the true dose-response model lies within the chosen model space, Wheeler and Bailer (2009) concluded that model averaging is superior to other commonly used approaches but may perform poorly otherwise and hence the suggestion that the model space includes a wide variety of model curvatures. The exercise of deciding on an appropriate model set should be guided by considerations of *parsimony* and *redundancy*. By parsimony, we mean balancing the number of candidate distributions with the variety of distributional shapes available. Redundancy considerations require us to avoid selecting distributions having similar shapes. This is important since the weighting mechanism of

Equation 9 will over-represent a particular SSD shape if two or more models fit the data equally well. To see this, consider three SSD models having *AIC* values of 1, 1, and 2 indicating the first two models fit the data equally well. Equation 9 assigns a weight of 0.384 to models 1 and 2 and a weight of 0.233 to model 3 meaning the single shape of the SSD represented by models 1 and 2 is given a combined weight of 0.768. Eliminating one of the redundant models from this calculation results in a down-weighting of the common shape represented by models 1 and 2 from 0.768 to 0.622 and a commensurate increase in the weight of model 3 from 0.233 to 0.378.

On balance, we believe model averaging provides a level of flexibility and parsimony that is difficult to achieve with a single SSD distribution. While subjective decisions still need to be made about the model set to which *AIC* weights are applied, guidelines and advice are available to assist the selection process.

Dealing with multimodality

Statistical mixture modelling (SMM): The technical challenge of multimodality was discussed earlier. An alternative strategy to data-splitting or weighting the lower portion of the SSD is to fit a *mixture* of statistical distributions to the complete toxicity data set. We refer to this as *statistical mixture modelling* (SMM)

In the remainder of this section, we outline how SMM may provide a way forward. Although the use of SMM has previously been used in ecotoxicology (e.g., Zajdlik et al. 2009; Zajdlik 2015), it has gained no traction with practitioners. This may be due to the lack of readily available software. In statistics, a mixture model is simply a weighted

combination of several individual probability models. Specifically, a statistical mixture $g(x;\Theta)$ of *k* distributions is:

$$g(x;\Theta) = \sum_{i=1}^{k} \lambda_i f(x;\theta_i) \quad 0 \le \lambda_i \le 1; \quad \sum_{i=1}^{k} \lambda_i = 1 \qquad (10)$$

where θ_i (possibly vector-valued) and λ_i are the parameter(s) and the weight associated with the i^{th} component distribution respectively and $\Theta = \bigcup_{i=1}^{k} \{\theta_i, \lambda_i\}$.

By way of example, consider the distribution of toxicity for a sample of heterotrophs and phototrophs shown in Figure 4. The bimodality for these data is clear. Using a single log-logistic distribution to model the pooled data gives an estimated HC5 of 1.09 compared to an estimated HC5 of 0.37 from a log-logistic fitted to just the phototrophs.

Instead of having to choose one or the other of these estimates, a mixture of 2 log-logistic distributions can be fitted to the pooled data. This requires the estimation of the 5 parameters in Equation 11.

$$g(x;\Theta) = \lambda f_1(x;\mu_1,\sigma_1) + (1-\lambda) f_2(x;\mu_2,\sigma_2) \qquad (11)$$

where $f_i(x;\mu_i,\sigma_i)$ is a log-logistic probability density function (*pdf*) and $\Theta^T = [\mu_1, \sigma_1, \mu_2, \sigma_2, \lambda]$. Here, μ_1 and σ_1 are the log shape and log scale parameters of the first log-logistic distribution, μ_2 and σ_2 are the log shape and log scale parameters of the second log-logistic distribution, and λ is a mixing parameter or weight to be applied to

the component distributions. The maximum likelihood estimate (mle) of Θ is $\hat{\Theta}^T = [9.143, 0.462, 2.267, 0.840, 0.584\ ]$.

Using Equation 11 with $\Theta = \hat{\Theta}$ we obtain an estimated HC5 of 1.81 which raises two important points: (i) the HC5 estimated from a mixture model is *not* equal to the weighted average of the individual HC5 values from the component distributions; and (ii) the estimated HC5 from a mixture model will lie between the HC5 computed using all the data and the HC5 using just the most sensitive species data. An indication of the adequacy of the fitted mixture model can be seen from Figure 5.

We believe statistical mixture modelling has an important role to play in ecotoxicology and, accordingly, we are currently developing code to incorporate this capability within the existing R-package ssdtools. While relatively parameter heavy (5 parameters for a mixture of two log-logistic distributions), statistical mixture models better match the inherent underlying functional process leading to bimodality in the first place compared to their univariate counterparts (e.g., Equation 1), i.e., they directly model bimodality as a mixture of two underlying univariate distributions that represent, for example, different modes of action. When using statistical mixture models within a model averaging approach, the high penalty in AICc associated with the increased number of parameters (*p* in Equation 7) when sample sizes are small will result in mixture models having low model weights when sample sizes are small and insufficient to support their robust estimation.

Left-tail weighting of the SSD: Some jurisdictions such as the US EPA exclusively use a small group of the most sensitive species when generating WQB values (Stephan et al.

1985). When using SSDs to derive WQBs, the question often arises of whether left tail (i.e., sensitive) species should have more weight when fitting the model and calculating HC5 values.

In practice, more weight could be given to the left-tail region by increasing the representation of sensitive species in the toxicity data set. When this is not a viable option, a relatively easy way to give greater weight to the toxicity data from the more sensitive taxa while still utilising *all* available data, is to augment the full data set with additional data resampled from the most sensitive species. There is, however, a large degree of subjectivity associated with this process – namely, deciding on a cut-off for and amount of additional weighting.

For example, if we add a copy of the toxicity data for the phototrophs in Figure 4 to the existing full data set and re-fit the log-logistic model, we obtain an estimated HC5 of 0.21. This highlights another difficulty – the estimated HC5 from this 'pseudo-sample' is smaller than the HC5 obtained from fitting an SSD to just the phototrophs.

The assumption of the statistical approach is that the species sensitivity data can be described by a single statistical distribution and that a model fit to all the data will provide the best estimate of an HC5.When this assumption is satisfied, there is no reason for giving extra weight to the left tail when fitting the SSD. However, because this assumption is invariably false, we may wish to improve the fit in the left portion of the distribution by down-weighting the influence of the extreme right tail observations. Such a procedure has been described by Liu et al. (2018). The concept is both simple and effective. Consider a random sample of *n* independent, identically distributed toxicity

values $\{X_1, X_2, ..., X_n\}$ from a family of distributions parameterized by the *k*-component vector Θ. The *k* values of Θ can be estimated by maximizing the censored likelihood given by Equation 12, where the largest *n-m* observations have been artificially censored.

$$L(\Theta; X_1, \cdots, X_n) = \left[1 - F_X\left(X_{(m)}; \Theta\right)\right]^{n-m} \prod_{i=1}^{m} f_X\left(X_{(i)}; \Theta\right) \quad (12)$$

where $X_{(i)}$, $i = 1, 2, ..., n$ are the order statistics and $f_X(\cdot; \Theta)$ and $F_X(\cdot; \Theta)$ are the probability density function and cumulative distribution function respectively.

Estimated HC5 values from log-logistic distributions obtained by maximizing Equation 12 for $m = \{17, ..., 7\}$ (corresponding to no censoring to right-censoring of all heterotroph data) were compared with estimated HC5 values obtained from fitting log-logistic distributions to only the non-censored portion of data (Table 2).

Although no general conclusions can be drawn from the results in Table 2, we see that generally differences between HC5 values from SSDs fitted using Equation 12 and SSDs fitted to only non-censored data are reasonably similar. More substantial differences arise as the level of censoring increases.

In summary, there are several potential options for dealing with multimodality: (i) use all data to fit the SSD using a unimodal model (i.e. do not account for bimodality); (ii) use only the data from the most sensitive species; (iii) use all data to fit the SSD using a statistical mixture model; and (iv) use all the data but assign greater weight to those values in the left-tail region (or alternatively, down-weight or censor more extreme values on the right).

The first option requires no subjective decisions other than those used in the planning and data collection stages of the SSD modelling exercise. However, this strategy may be problematic when the fit in the left-tail appears worse than the fit in the middle and upper

regions of the SSD. Although the second strategy is appealing and is consistent with WQB derivation and risk assessment methodologies in Australia/New Zealand and elsewhere, it is not unequivocal when there is overlap in the bimodal distributions. As our limited analyses show, the high degree of subjectivity associated with left-tail weighting and the resultant impact on estimated HC5 values would suggest this approach (option 4) is sub-optimal. Like the first option, the third option of fitting a statistical mixture model requires no additional subjective decisions. It enjoys the same advantage of using all the data, but unlike other options it does so in a way that attempts to provide an equally good fit in all regions of the SSD. For this reason, we suggest that statistical mixture models be considered for modelling bimodal distributions, whilst recognising that, in some cases (e.g. chemical-specific risk assessments with substances of known mode of action), it may be more appropriate to split the data and derive taxa-specific HCx estimates.

SOFTWARE DEVELOPMENT

Computations associated with fitting and using SSDs are invariably complex and best handled by purpose-built software such as those listed in Table 1. Some of these software tools have been in existence for over 20 years and are both used and endorsed by regulatory agencies for the purpose of setting WQBs for marine and freshwater systems. It is not our intention to provide a comprehensive review of all these tools, but rather to highlight new additions and features. Accordingly, we focus on two products: the ssdtools R package (Thorley and Schwarz 2018) together with the associated shinyssdtools app (Dalgarno 2018); and the recently released SSD Toolbox (Center for Computational Toxicology and Exposure 2020).

Shinyssdtools app

ssdtools is an R software package developed for the British Columbia Ministry of Environment and Climate Change Strategy (Thorley and Schwarz 2018). shinyssdtools is a web-based graphical user interface to ssdtools which uses the R shiny package (Chang et al. 2019).

Web deployment of apps is becoming increasingly popular and has several advantages over standalone software. In particular, the user is guaranteed of using the most up-to-date version of the software as well as being able to run analyses from any device that supports browsing. Furthermore, being an R package means the ssdtools source code is completely transparent and available for local modification. As noted in the Current status section, issues such as statistical consistency and transparency need to be considered when using SSDs for various purposes, and there is likely to be demand for both modifiable and "locked" (i.e., compiled) code.

shinyssdtools is currently hosted at https://bcgov-env.shinyapps.io/ssdtools although shinyssdtools is itself an R package (https://github.com/bcgov/shinyssdtools) that can be run locally. The interface is clean and simple and allows the user to either cut and paste data directly into the app or upload from a local csv file. Although individual distributions can be used to obtain HCx values, the focus and strength of ssdtools is its intrinsic use of model-averaging. The R package ssdtools and the accompanying Shiny app (Dalgarno 2018) currently fits the log-normal, log-logistic and gamma by default and optionally, the log-Gumbel, Gompertz and Weibull. The default distributions were selected in accordance with our concepts of *parsimony* and *redundancy*.

The log-normal distribution was selected as the starting distribution given the data are for effect concentrations. The log-normal distribution does have a couple of characteristics that need to be considered when fitting species sensitivity data. First, on the logarithmic scale, the normal distribution is symmetrical and there are no a priori grounds on which

to make any assumption about an SSDs shape or scale whether that be on the original or log-transformed scale. Second, the log-normal distribution decays quickly in the tails giving narrow tails that may not adequately fit the data.

The log-logistic distribution was selected as it is often used as a candidate SSD primarily because of its analytic tractability (Aldenderg and Slob 1993). It was included because it has wider tails than the log-normal and because it is a specific case of the more general Burr family of distributions (Burr 1942, Shao 2000).

The gamma distribution is a two-parameter distribution commonly used to model failure times or time to events. For use in modelling species sensitivity data, the gamma distribution has two key features that provide additional flexibility when added to the log-normal distribution: (i) it is asymmetrical on the logarithmic scale; and (ii) it has wider tails. The Weibull distribution was also considered as a default distribution, but the gamma distribution is generally more flexible whilst capturing similar shaped distributions to the Weibull.

SSD Toolbox

The SSD Toolbox is a US Environmental Protection Agency product. It is made available as a Windows executable file and can be downloaded from https://epa.figshare.com/articles/Species_Sensitivity_Distribution_SSD_Toolbox/11971392

Before using SSD Toolbox, the user must also download and install version 9.5 of the MATLAB® Runtime Compiler (MCR) from Mathworks. The MCR software enables the

compiled code to execute without having to purchase the MATLAB® product. It is however a resource-hungry piece of software with its 88,000+ files consuming 3.75GB of hard disk space.

Overall, SSD Toolbox is a comprehensive piece of software that essentially performs the same functions as ssdtools with some additional features (Table 1). It has a GUI that is adequate, but not as aesthetically appealing as the shinyssdtools app. There are 6 theoretical distributions for SSD fitting log-transformed data (normal; logistic; triangular; Gumbel; Weibull; Burr) using up to 4 fitting methods (maximum likelihood; moment matching; *cdf* linearization; and Bayesian methods). Although the triangular distribution is formally used by the USEPA for deriving ambient water quality criteria, this distribution is a curious inclusion given that it has tail characteristics that are not generally encountered in practice and therefore not widely used as a realistic SSD. The *cdf* linearization method is also an unusual choice as this is a relatively crude way of fitting distributions and provides SSD parameter estimates that do not necessarily share desirable statistical properties enjoyed by other methodologies such as maximum likelihood estimation.

FUTURE DIRECTIONS

Limitations and conceptual difficulties with SSD modelling were acknowledged in the introduction. Despite these and acknowledging a potential future for non-SSD (i.e., distribution-free) methods, we are of the view that the SSD methodology remains the most credible and statistically defensible way of establishing protective concentrations of toxicants in aquatic environments in the short to medium term. The methodological

developments described in the current paper have addressed some long-standing issues such as choice of an appropriate probability model and difficulties introduced by bi- and multi-modality, while numerous refinements to other aspects of SSDs have been published by others over the past 20 years (see references cited in the Introduction). Nevertheless, there are still several long-standing and unresolved issues with SSDs, including small sample bias in SSD parameter estimates, convergence issues with more complex models and other issues as identified by Belanger et al. (2017). Moreover, there are other areas where progress has been made and/or further investigation may be warranted (e.g., use of censored data (Aldenberg 2015) and Bayesian methods (Fox 2010; Takehiko and Kashiwagi 2010)). In terms of our own R&D efforts, we have identified the following priority issues that will form the basis of further collaboration between Australian and Canadian jurisdictions:

Numerical stability issues

The use of the Burr family of distributions has been central to the derivation of GVs in Australia and New Zealand for over 20 years. While offering a high degree of flexibility, experience with these distributions during that time has repeatedly highlighted numerical stability and convergence issues when estimating parameters using maximum likelihood. This is thought to be due to the high degree of collinearity between parameter estimates and/or relatively flat likelihood profiles. Companion issues to be explored during this phase include estimation strategies and identification of initial values for iterative methods.

Benchmark data sets

Lack of agreement in derived quantities such as an HCx arising from different SSD modelling strategies and tool development undermines the credibility of the methodology. As argued in this paper, we believe it is both desirable and necessary to assemble a collection of reference data sets having certified properties that can be used to evaluate SSD methodologies and software tools. We envisage that this collection will be comprised of both real and synthetic (i.e. computer-generated) data sets and will have the ability to test both the accuracy and stability of SSD software.

Statistical mixture modelling

We propose to continue development and refinement of statistical mixture modelling (SMM) methodologies. This includes 1) identifying optimal parameter estimation strategies, and 2) an assessment of the performance of AICc-based model weighted averaging using candidate model sets that include mixture distributions. If the approach proves robust across a range of sample sizes and scenarios, it may be possible to incorporate mixture distributions as an option within the ssdtools package and shiny app.

HCx and CI estimation

Further work is required to understand the strengths and weaknesses of competing methods of estimating HCx values and associated confidence intervals post-distribution fitting. The software tools listed in Table 1 employ a mixture of strategies including: inversion of the fitted *cdf*, bootstrapping, and numerical approximations such as the delta

method. We also plan to investigate the potential of profile-likelihood based confidence intervals as a more robust and defensible strategy.

Identification of default distributions for model averaging

We strongly support model averaging as a means of (partially) resolving the issue of distribution selection in SSD modelling. How to make a rational and defensible selection of the default set of distributions to be used in the model averaging is an open issue. After all, model averaging can only assign weights to distributions in the candidate list and is blind to potentially better, but unspecified alternatives.

In conclusion, we note that advances in software architectures have opened new possibilities for researchers and practitioners to interact and contribute to SSD tool development in ways that hitherto have not been possible. The challenge as we see it now is how to better coordinate these interactions and avoid unnecessary duplication of effort and software redundancy. To this end, our participation in an SSD modelling workshop in Victoria, BC, in November 2019 was the first tentative step towards jurisdictional harmonization of methodological approaches and SSD tool development for Australia/New Zealand and Canada.

Given the different policy objectives, levels of risk tolerance, and species compositions, global harmonization of WQB derivation methods may be difficult to achieve, however, specific aspects of the WQB derivation method can be standardized to improve the comparison of WQBs across jurisdictions, to increase collaboration and to reduce duplication of effort. For example, Warne et al. (2018) called for increased effort to harmonize data assessment procedures such that jurisdictions can access a common

database of toxicity data for WQB derivation. Sharing data sets would greatly reduce the effort undertaken by individual jurisdictions when deriving WQBs and would also remove this source of variability when comparing WQBs across jurisdictions. Pursuing a goal of international harmonization of key aspects of WQB derivation would be greatly assisted by a more formal and regularly convened multi-national group of experts that considers and investigates opportunities for improvements to and harmonization of WQB derivation methods, and which makes recommendations (based on research and development) that jurisdictions could then adopt as their respective timelines allow.

Although there have been various fora aimed at advancing the science of WQB derivation (e.g., ECETOC, and the "Environmental Quality Standards for Protection of the Aquatic Environment" (EQSPAE) series of conferences), one-off or even periodic efforts will not be as effective at enabling long-term material advances in the way that WQBs are derived at an international scale. While the specific statistical issues described in the current paper will be pursued through the current Australia/New Zealand and Canada collaboration, a broader multi-national forum is needed to evaluate new and improved approaches to WQB derivation and facilitate their common adoption.

Data Availability Statement—Data and R code are available from the corresponding author

ACKNOWLEDGEMENTS

The authors are most grateful to the anonymous reviewers for their many useful suggestions that greatly improved the organisation and clarity of the final manuscript.

REFERENCES

Akaike H. 1973. Information theory and the maximum likelihood principle. In: Petrov BN, Coaki F, eds, *2nd International Symposium on Information Theory,* Akademiai Kiado, Budapest, Hungary.

Aldenberg T. 2015. In response: Challenges when weighing evidence about environmental risks - an industry perspective. *Environ Toxicol Chem* 34:2442-2444.

Aldenberg T, Slob W. 1993. Confidence limits for hazardous concentrations based on logistically distributed NOEC toxicity data. *Ecotox Environ Saf* 25:48-63.

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

ANZG 2018. Australian and New Zealand guidelines for fresh and marine water quality. Australian and New Zealand Governments and Australian state and territory governments, www.waterquality.gov.au/anzguidelines.

Barry S, Henderson B. 2014. Burrlioz 2.0, Commonwealth Science and Industrial Research Organisation, Canberra, Australia, accessed 24 December, 2014.

Batley GE, van Dam RA, Warne MStJ, 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.

Belanger S, Barron M, Craig P, Dyer S, Galay-Burgos M, Hamer M, Marshall S, Posthuma L, Raimondo S, Whitehouse P. 2017. Future needs and recommendations in the development of species sensitivity distributions: estimating toxicity thresholds for aquatic ecological communities and assessing impacts of chemical exposures. *Integr Environ Assess Manag* 13:664-674.

Belanger SE, Carr GJ. 2019. SSDs revisited: part II—practical considerations in the development and use of application factors applied to species sensitivity distributions. *Environ Toxicol Chem* 38:1526-1541.

British Columbia Ministry of Environment and Climate Change Strategy, 2019. Derivation of Water Quality Guidelines for the Protection of Aquatic Life in British Columbia. Water Quality Guideline Series, WQG-06. Prov. B.C., Victoria B.C., Canada, 50 pp.

Burnham KP, Anderson DR. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, Second Edition, Springer-Verlag New York.

Burr IW. 1942. Cumulative frequency functions. *Annal Math Stat* 13:215–232.

Campbell E, Palmer M, Shao Q, Warne M, Wilson D. 2000. Burrlioz: A Flexible Approach to Species Protection, TIES/SPRUCE 2000 Conference, 4–8 September 2000, University of Sheffield, UK.

Carr GJ, Belanger SE. 2019. SSDs revisited: Part I—A framework for sample size guidance on species sensitivity distribution analysis. *Environ Toxicol Chem* 38:1514–1525.

CCME 2007. A Protocol for the Derivation of Water Quality Guidelines for the Protection of Aquatic Life. Canadian environmental quality guidelines, 1999. Canadian Council of Ministers of the Environment, Winnipeg, MB, Canada. Available at: https://www.ccme.ca/files/Resources/supporting_scientific_documents/protocol_aql_2007e.pdf.

CCME 2009. Canadian water quality guidelines for the protection of aquatic life: Boron. Canadian environmental quality guidelines, 1999. Canadian Council of the Ministers of the Environment, Winnipeg, MB, Canada. Available at: http://ceqg-rcqe.ccme.ca/download/en/324/

CCME 2013. Determination of hazardous concentrations with species sensitivity distributions, SSD Master. Canadian Council of Ministers of the Environment Report 38, Ottawa, ON, Canada.

CCME 2015. Canadian water quality guidelines for the protection of aquatic life: Silver. Canadian environmental quality guidelines, 1999. Canadian Council of the Ministers of the Environment, Winnipeg, MB, Canada. Available at: http://ceqg-rcqe.ccme.ca/download/en/355/

CCME 2019. Scientific criteria document for the development of the Canadian water quality guidelines for the protection of aquatic life: Manganese. Canadian Council of Ministers of the Environment, Winnipeg, MB, Canada.. Available at: https://www.ccme.ca/files/Resources/supporting_scientific_documents/Manganese%20CWQG%20SCD%20EN%20(secured).pdf

CCME 2020. Draft scientific criteria document for the development of the Canadian water quality guidelines for the protection of aquatic life: neonicotinoid insecticides. Canadian Council of Ministers of the Environment, Winnipeg, MB, Canada. Available at: https://www.ccme.ca/files/DRAFT%20CWQG%20NNIs%20SCD%20EN%20v3.5%20secure.pdf

Center for Computational Toxicology and Exposure E. 2020. Species Sensitivity Distribution (SSD) Toolbox. doi:10.23645/epacomptox.11971392.v1.

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. 2019. shiny: Web Application Framework for R. https://CRAN.R-project.org/package=shiny

Chapman PF, Reed M, Hart A, Roelofs W, Aldenberg T, Solomon K, Tarazona J, Liess M, Byrne P, Powley W, Green J, Ferson S, Galicia H. 2007. Methods of uncertainty analysis. In: Hart A, ed, EUFRAM *Concerted Action to Develop a European Framework for Probabilistic Risk Assessment of the Environmental Impacts of Pesticides, Volume 2, Detailed Reports on Role, Methods, Reporting and Validation.* Available at https://cordis.europa.eu/project/id/QLK5-CT-2002-01346 Accessed 20/8/2019.

Charles S, Veber P, Delignette-Muller ML. 2018. MOSAIC: a web-interface for statistical analyses in ecotoxicology. *Environ Sci Pollut Res Int* 12:11295-11302. doi: 10.1007/s11356-017-9809-4.

Craig, PS. 2013. Exploring novel ways of using species sensitivity distributions to establish PNECs for industrial chemicals: Final report to Project Steering Group 3 April 2013. Technical Report, DU. Available from http://dro.dur.ac.uk/13383/.

Dalgarno S. 2018. ssdtools: A shiny web app to analyse species sensitivity distributions Prepared by Poisson Consulting for the Ministry of the Environment, Victoria, British Columbia, Canada. Available at https://bcgov-env.shinyapps.io/ssdtools. Accessed 29/5/2020.

D'Andrea MF, Brodeur JC. 2019. shinyssd v1.0: Species sensitivity distributions for ecotoxicological risk assessment. JOSS 4:785. doi:10.21105/joss.00785.

De Zwart D, Posthuma L. 2005. Complex mixture toxicity for single and multiple species: proposed methodologies. *Environ Toxicol Chem* 24:2665–2676.

Dormann CF, Calabrese JM, Guillera-Arroita G, Matechou E, Bahn V, Barton K, Beale CM, Cuiti S, Elith J, Gerstner K, Gelat J, Keil P, Lahoz-Monfort JJ, Pollock LJ, Reineking B, Roberts DR, Schroder B, Thuiller W, Warton DI, Wintle BA, Wood SN, Woest RO, Hartig F. 2018. Model averaging in ecology: a review of Bayesian, information- theoretic, and tactical approaches for predictive inference. *Ecol Monogr* 88:485-504.

Dyer SD, Versteeg DJ, Belanger SE, Chaney JG, Mayer FL. 2006. Interspecies correlation estimates (ICE) predict protective environmental concentrations. *Environ. Sci. Technol.* 40:3102–3111.

ECETOC 2014. Estimating toxicity thresholds for aquatic ecological communities from sensitivity distributions. European Centre for Ecotoxicology and Toxicology of Chemicals. Workshop Report 28, Brussels, Belgium, 98 pp.

Efron B, Tibshirani R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci* 1:54-75.

Etterson M. 2020. User's Manual: SSD Toolbox Version 1.0. US Environmental Protection Agency, Office of Research and Development, Center for Computational Toxicology and Exposure.

European Commission 2011. Guidance Document No. 27 Technical Guidance for Deriving Environmental Quality Standards: Common Implementation Strategy for the Water Framework Directive (2000/60/EC) Technical Report - 2011 – 055. Available at https://circabc.europa.eu/sd/a/0cc3581b-5f65-4b6f-91c6-433a1e947838/TGD-EQS%20CIS-WFD%2027%20EC%202011.pdf, accessed 20/8/2019

Fisher R, van Dam RA, Batley GE, Fox DR, Harford AJ, Humphrey CL, King CK, Menendez P, Negri AP, Proctor A, Shao Q, Stauber JL, van Dam JW, Warne MStJ. 2019. Key Issues in the Derivation of Water Quality Guideline Values: a Workshop

Report. Australian Institute of Marine Science Report, Crawley, WA, Australia, 57 pp. DOI:10.13140/RG.2.2.29774.82241

Forbes VE, Calow P. 2002. Species sensitivity distributions revisited: a critical appraisal. *Hum Ecol Risk Assess* 8:473–492.

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

Fox DR. 2010 A Bayesian approach for determining the no effect concentration and hazardous concentration in ecotoxicology. *Ecotox Environ Saf* 73:123–131.

Fox DR. 2016. Contemporary methods for statistical design and analysis. In: Blasco J, Chapman P, Campana O, Hampel M, eds, *Marine Ecotoxicology: Current Knowledge and Future Issues*, Elsevier, San Diego, CA, USA, pp. 35 -70.

Freeman JB, Dale R. 2013. Assessing bimodality to detect the presence of a dual cognitive process. *Behav Res Meth* 45:83–97.

Gelman A, Lee, D, Guo, J. 2015. Stan: A probabilistic programming language for Bayesian inference and optimization. *J Educ Behav Stat* 40:530-543. doi:10.3102/1076998615606113.

Hickey GL, Craig PS. 2012. Competing statistical methods for the fitting of normal species sensitivity distributions: recommendations for practitioners. *Risk Anal* 32:1232–1243.

Intrinsik Environmental Sciences Inc. 2013. Determination of hazardous concentrations with species sensitivity distributions. SSD MASTER Version 3.0. Report Prepared for the Canadian Council of Ministers of the Environment.

Kon Kam King G, Veber P, Charles S, Delignette-Muller ML. 2014. MOSAIC SSD: a new web tool for species sensitivity distribution to include censored data by maximum likelihood. *Environ Toxicol Chem* 33:2133–2139.

Kooijman SALM. 1987. A safety factor for LC50 values allowing for differences in sensitivity among species. *Water Res* 21:209-221.

Kullback S. 1959. Information Theory. Wiley, New York, USA.

Lepper P. 2005. Manual on the methodological framework to derive environmental quality standards for priority substances in accordance with Article 16 of the Water Framework Directive (2000/60/EC). Fraunhofer-Institut Molecular Biology and Applied Ecology Report, Schmallenberg, Germany, 47 pp.

Liu Y, Salibián- Barrera M, Zamar RH, Zidek JV. 2018. Using artificial censoring to improve extreme tail quantile estimates. *J Royal Stat Soc* C 67:791-812.

Longley JW. 1967. An appraisal of least squares programs for the electronic computer from the point of view of the user. *J Amer Stat Assoc* 62:819-841.

Mancini D, Lardo A, De Angelis M. 2020. Efforts towards openness and transparency of data: a focus on open science platforms. In: Lazazzara A, Ricciardi F, Za S, eds, *Exploring Digital Ecosystems*, Springer, Cham, Switzerland, pp. 67-84.

Moore DR, Priest CD, Galic N, Brain RA, Rodney SI. 2019. Correcting for phylogeneticautocorrelation in species sensitivity distributions. *Integr Environ Assess Manag* 16:53-65. doi:10.1002/ieam.4207.Munafò MR, Nosek BA, Bishop DV, Button KS, Chambers CD, Du Sert NP, Simonsohn U, Wagenmakers EJ, Ware JJ, Ioannidis JP. 2017. A manifesto for reproducible science. *Nat Hum Behav* 1:1-9.

Newman MC, Ownby DR, Mezin 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. *Environ Toxicol Chem* 19:508–515.

OECD 1992. Report of the OECD Workshop on extrapolation of laboratory aquatic toxicity data to the real environment. OECD Environment Monograph No. 59, Organisation for Economic Cooperation and Development, Paris, France.

Pfister R, Schwarz KA, Janczyk M, Dale R, Freeman JB. 2013. Good things peak in pairs: a note on the bimodality coefficient. *Front Psychol* 4:700.

Plummer M. 2003. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. Proceedings of the third international workshop on distributed statistical computing (DSC 2003), Vienna, Austria, 8 pp. http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Drafts/Plummer.pdf

Posthuma L, Suter GW, Traas TP. 2002. Species Sensitivity Distributions in Ecotoxicology. CRC Press, Boca Raton, FL, USA, 616 pp.

Posthuma L, van Gils J, Zijp MC, van de Meent D, de Zwart D. 2019. Species sensitivity distributions for use in environmental protection, assessment, and management of aquatic ecosystems for 12 386 chemicals. *Environ Toxicol Chem* 38:905–917.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schwarz CJ, Tillmanns AR. 2019. Improving statistical methods to derive species sensitivity distributions. Water Science Series, WSS2019-07, Province of British Columbia, Victoria, B.C., Canada, 29 pp.

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

Smith EP, Cairns J. 1993. Extrapolation methods for setting ecological standards for water quality: statistical and ecological concerns. *Ecotoxicology* 2:203–219.

Stephan CE. 2002. Use of species sensitivity distributions in the derivation of water quality criteria by the U.S. Environmental Protection Agency. In: Posthuma L, Suter GW, Traas TP, eds, Species Sensitivity Distributions in Ecotoxicology, CRC Press, Boca Raton, FL, USA, pp. 211-220.

Stephan CE, Mount DI, Hansen DJ, Gentile JH, Chapman GA, Brungs WA. 1985. Guidelines for Deriving Numerical National Water Quality Criteria for the Protection of Aquatic Organisms and their Uses. U.S. Environmental Protection Agency Report, Office of Research and Development, Washington, DC, USA, 59 pp.

Suter GW. 2002. North American history of species sensitivity distributions. In: Posthuma L, Suter GW, Traas TP, eds, Species Sensitivity Distributions in Ecotoxicology, CRC Press, Boca Raton, FL, USA, pp. 11-18.

Tadikamalla PR. 1980. A look at Burr and related distributions. *Int Stat Rev* 48: 337-344.

Takehiko IH, Kashiwagi N. 2010. A Bayesian method for deriving species-sensitivity distributions: selecting the best-fit tolerance distributions of taxonomic groups. *Hum Ecol Risk Assess* 16:251-263.

Thorley J, Schwarz C. 2018. ssdtools: An R package to fit species sensitivity distributions. J *Open Source Softw* 3:1082. doi: 10.21105/joss.01082.

USEPA. 2004. SSD Generator. U.S. Environmental Protection Agency, Center for Public Health and Environmental Assessment (CPHEA), Washington, DC, USA. Available at: https://www.epa.gov/caddis-vol4/caddis-volume-4-data-analysis-download-software

Van Der Hoeven N. 2001. Estimating the 5-Percentile of the species sensitivity distributions without any assumptions about the distribution. *Ecotoxicology* 10:25-34.

van Straalen NM, Denneman CAJ. 1989. Ecotoxicological evaluation of soil quality criteria. *Ecotox Environ Saf* 18:241-251.

Van Vlaardingen PLA, Traas TP, Wintersen AM, Aldenberg T. 2004. ETX 2.0. A program to calculate hazardous concentrations and fraction affected, based on

normally distributed toxicity dataNational Institute for Public Health and the Environment (RIVM). Report No. 601501028/2004, Bilthoven, the Netherlands, 68 pp.

Wang Y, Zhang L, Meng F, Zhou Y, Jin X, Giesy JP, Liu F. 2015. Improvement on species sensitivity distribution methods for deriving site-specific water quality criteria. *Environ Sci Pollut Res* 22:5271–5282.

Warne MStJ. 1998. Critical review of methods to derive water quality guidelines for toxicants and a proposal for a new framework. Supervising Scientist Report 135, Supervising Scientist, Canberra, ACT, Australia, 82 pp.

Warne MStJ, 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.

Wheeler, MW, Bailer, AJ. 2009. Comparing model averaging with other model selection strategies for benchmark dose estimation. *Environ Ecol Stat* 16:37-51.

Wheeler JP, Grist EPM, Leung KMY, Morritt D, Crane M. 2002a. Species sensitivity distributions: data and model choice. *Mar Pollut Bull* 45:192–202.

Wheeler JR, Leung KMY, Morritt D, Whitehouse P, Sorokin N, Toy R, Holt M, Crane M. 2002b. Freshwater to saltwater toxicity extrapolation using species sensitivity distributions. *Environ Toxicol Chem* 21:2459–2467.

Zajdlik B. 2005. Statistical analysis of the SSD approach for development of Canadian water quality guidelines. Report for CCME Project Number 354-2005, Zajdlik and Associates Inc, Rookwood Ottawa, Ontario, Canada, 40 pp.

Zajdlik B. 2006. Potential statistical models for describing species sensitivity distributions. Report for CCME Project Number 382-2006, Zajdlik and Associates Inc, Rookwood, Ontario, Canada, 25 pp.

Zajdlik B. 2015.The Statistical Derivation of Environmental Quality Guidelines. PhD. Thesis, University of Waterloo, Waterloo, Ontario, Canada.

Zajdlik BA, Dixon DG, Stephenson G. 2009. Estimating water quality guidelines for environmental contaminants using multimodal species sensitivity distributions: a case study with atrazine. *Hum Ecol Risk Assess* 15:554–564.

Zhang J, Bailer JA, Oris JT. 2012. Bayesian approach to estimating reproductive inhibition potency in aquatic toxicity testing. *Environ Toxicol Chem* 31:916-927.

FIGURES

Figure 1. Q-Q plots and goodness-of-fit test statistics for four probability distributions fitted to a small toxicity data set (AD refers to the Anderson-Darling goodness-of-fit statistic)

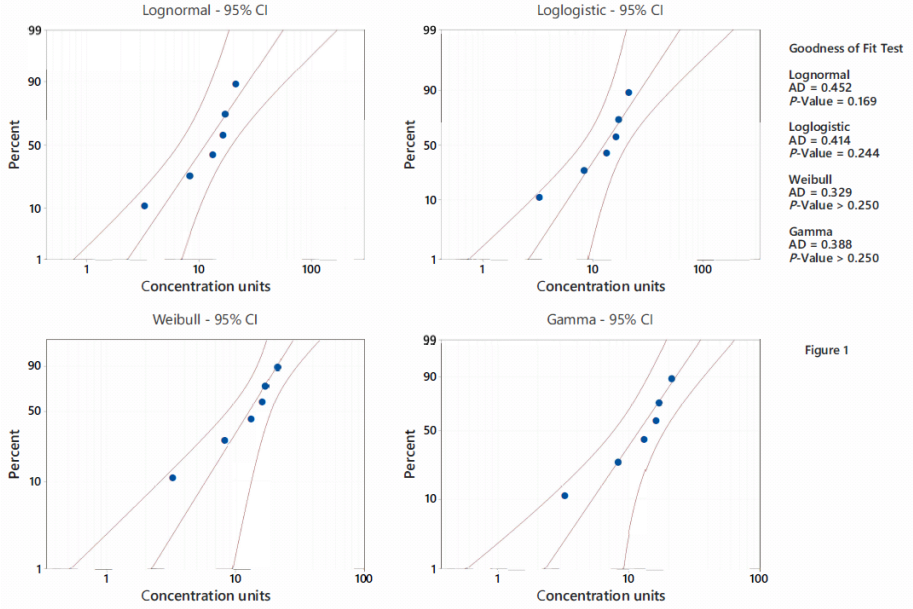

Figure 2. Comparison of probability density functions for triangular (blue curve) and lognormal (red curve) distributions. Full view (a) and left tail-view (b).

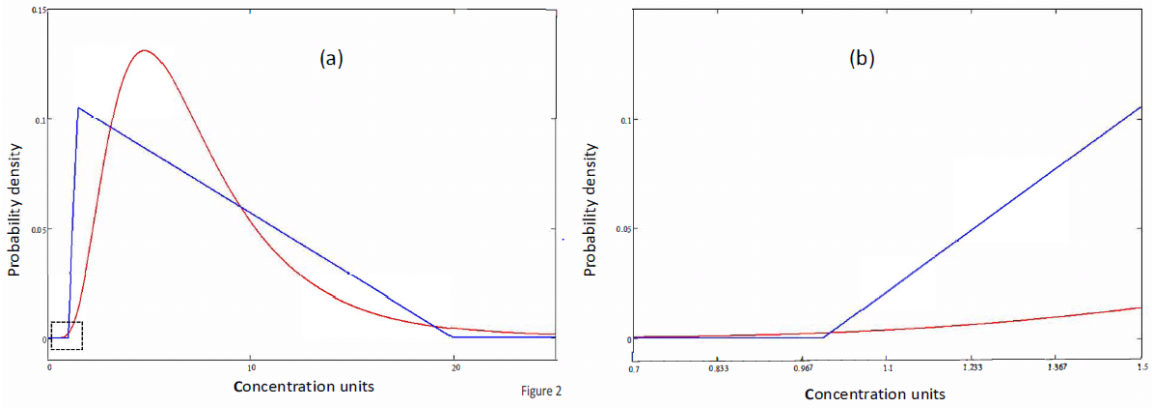

Figure 3. Bimodal distribution given by Equation 9 with μ=2 and σ=1.75. Modal values at x=0.262 and x=3.646.

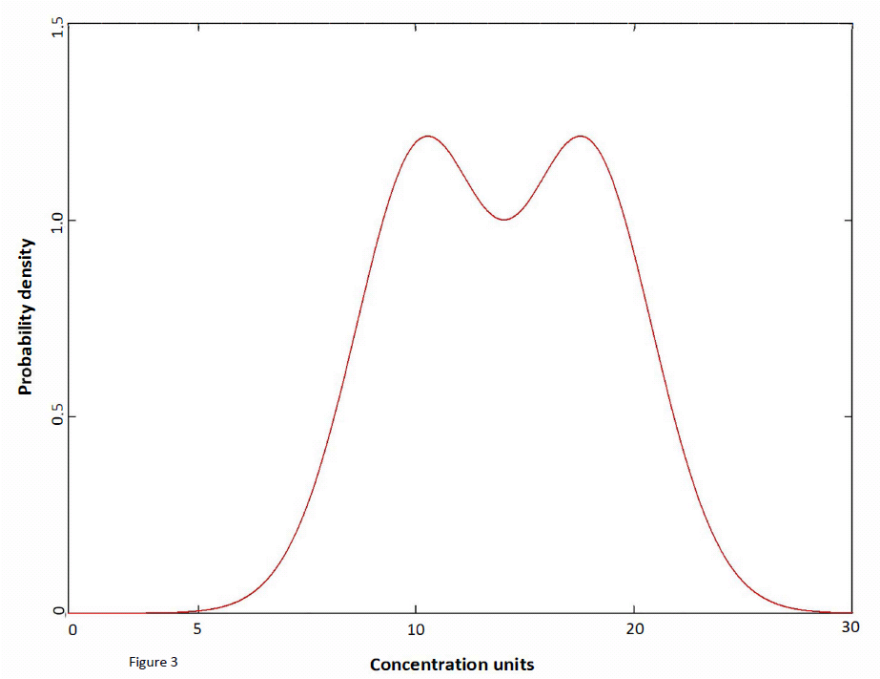

Figure 4. Empirical SSD (bars) for data comprised of 10 heterotrophs and 7 phototrophs with smooth overlaid (solid lines)

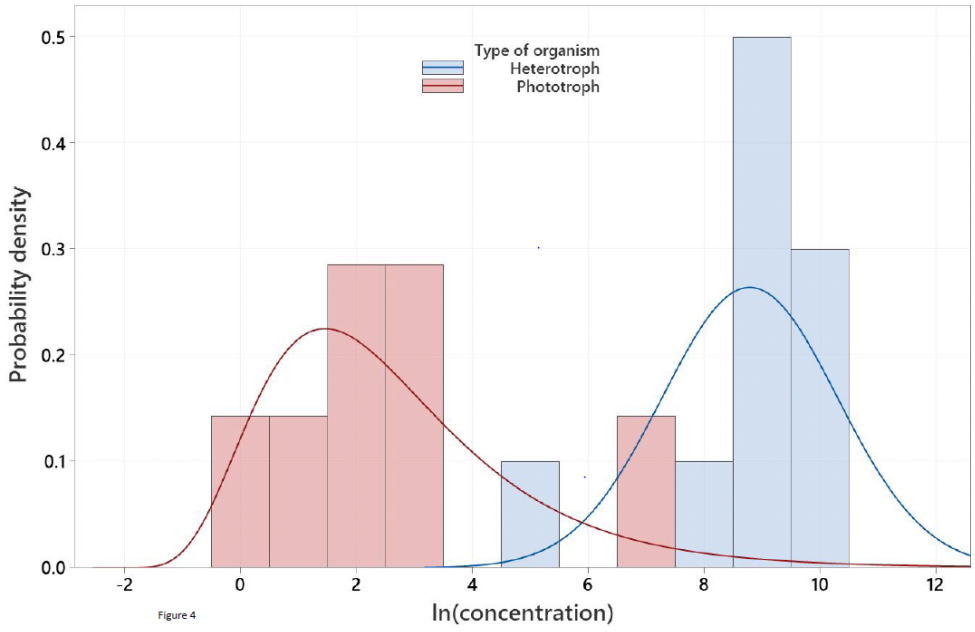

Figure 5. Empirical probability distribution (a) and cumulative probability function (b) for data in Figure 4 together with fitted mixture of 2 log-logistic distributions (solid blue line)

Figure 5

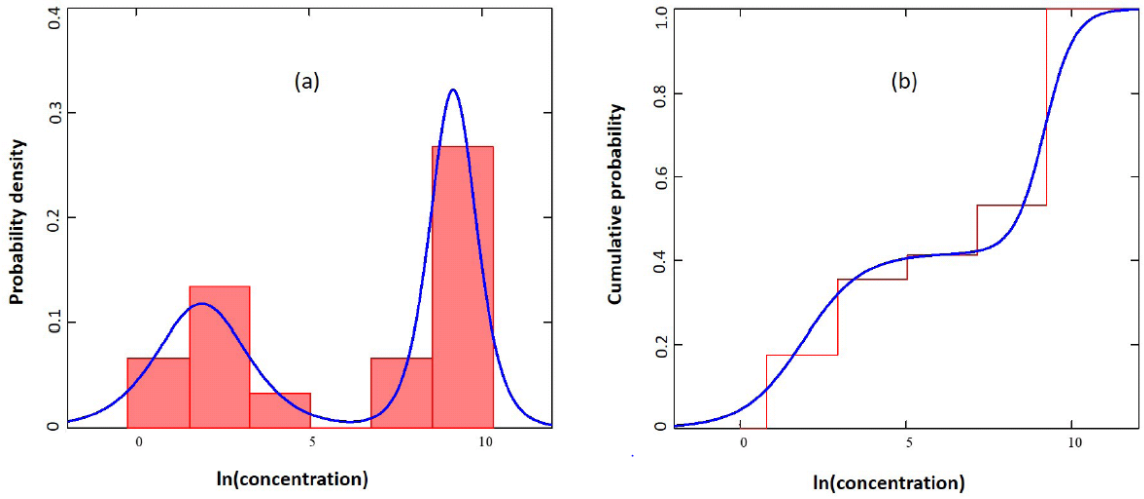

Table 1. Software tools to fit distributions to data using Least Squares (LS), Moment Matching (MM), Maximum Likelihood (ML), CDF linearization (CL), and Bayesian (BY) analysis.

Software		**Burrlioz**	**ETX 2.0**	**hSSD**
Current Version		2	2.2	
Analytic Method		ML	LS, ML	BY
Distributions	log-logistic	x		
	log-normal		x	
	Weibull			
	log-Gumbel			
	burr III	x		
	log-t			x

	Pareto			
	log-triangular			
	gamma			
	Gompertz			
Functionality	Model Averaging			
	Censored Data			x
	Hierarchical			taxonomic
	HCx	1, 5, 10, 20	5,50	1,2,…,98,99
	Confidence Intervals	yes	yes	MCMC
	Computer Languages	R	Visual Basic	MATLAB
Platform	*linux*			*x*
	mac			*x*
	windows	*x*	*x*	*x*
Other details	*GUI*	*x*	*x*	*x*
	Code/Web Interface/URL	*https://research.csiro.au/software/burrlioz/*	*https://rvs.rivm.nl/risicobeoordeling/modellen-voor-risicobeoordeling/ETX*	*http://www.ecetoc.org/tools/ tool/*
	Country	*Australia, NZ*	*Netherlands*	*England*
	Reference	*Campbell et al. (2000); Barry and*	*Van Vlaardingen et al. (2004)*	*Craig (2013)*

		Henderson (2014)		

Table 2. Comparison of HC5 estimates from log-logistic SSD as a function of censoring of largest i observations

	Estimated HC_5	
Largest *n-m* observations censored	**Equation 12**	**SSD fitted to non-censored data**
0	1.086	1.086
1	0.878	0.823
2	0.665	0.604
3	0.522	0.453
4	0.393	0.354
5	0.316	0.275
6	0.228	0.233
7	0.144	0.226
8	0.161	0.288
9	0.173	0.443
10	0.858	0.858