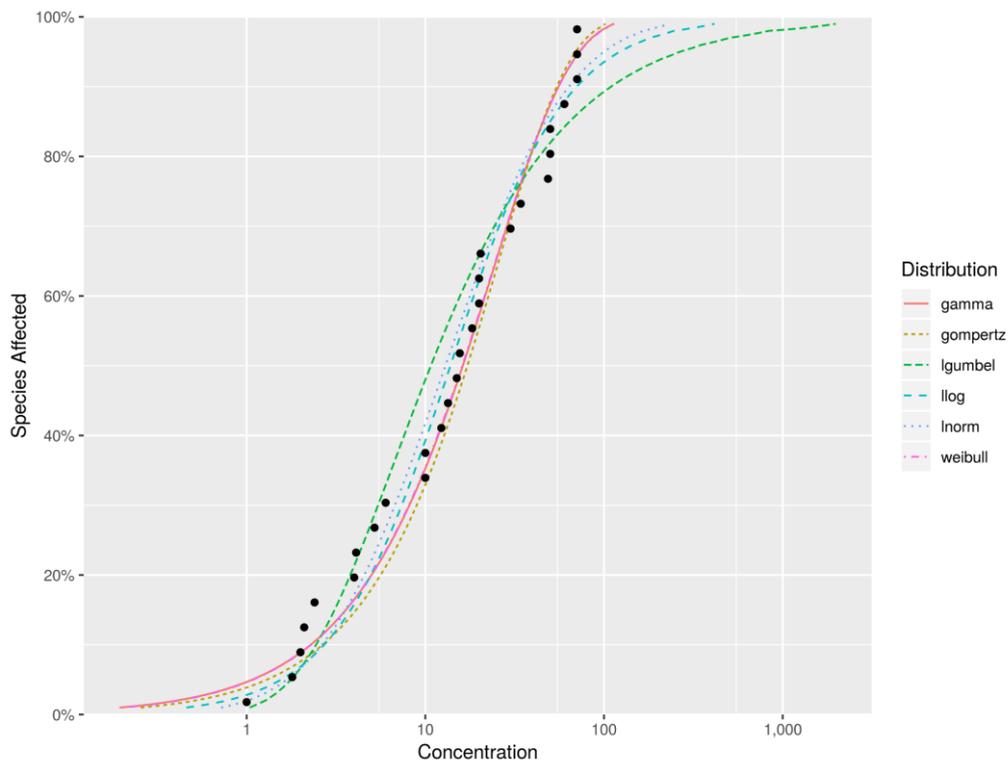


Improving Statistical Methods for Modeling Species Sensitivity Distributions

Carl Schwarz and Angeline Tillmanns



July 2019



The **Water Science Series** are scientific technical reports relating to the understanding and management of B.C.'s water resources. The series communicates scientific knowledge gained through water science programs across B.C. government, as well as scientific partners working in collaboration with provincial staff. For additional information visit: <http://www2.gov.bc.ca/gov/content/environment/air-land-water/water/water-science-data/water-science-series>.

ISBN: 978-0-7726-7773-0

Citation:

Schwarz, C.J. and A.R. Tillmanns. 2019. Improving statistical methods to derive species sensitivity distributions. Water Science Series, WSS2019-07, Province of British Columbia, Victoria.

Author's Affiliation:

Carl J. Schwarz, Ph.D.,P.Stat.
StatMath Comp Consulting by Schwarz
cschwarz.stat.sfu.ca@gmail.com

Angeline R. Tillmanns, Ph.D.,R.P.Bio
BC Ministry of Environment and Climate Change Strategy
Angeline.Tillmanns@gov.bc.ca

© Copyright 2019

Acknowledgements:

The authors gratefully acknowledge reviewers from Environment and Climate Change Canada (ECCC) and members of the Guideline Project Team of the Canadian Council for Ministers of the Environment for their helpful comments. We would also like to extend a special thank-you to Kathleen MacTavish at ECCC for her contributions to the discussion on model averaging.

Disclaimer: The use of any trade, firm, or corporation names in this publication is for the information and convenience of the reader. Such use does not constitute an official endorsement or approval by the Government of British Columbia of any product or service to the exclusion of any others that may also be suitable. Contents of this report are presented for discussion purposes only. Funding assistance does not imply endorsement of any statements or information contained herein by the Government of British Columbia.

EXECUTIVE SUMMARY

Water quality guidelines (WQGs) are used as benchmarks to interpret environmental risks related to changes in water quality. In B.C., the Ministry of Environment and Climate Change Strategy (ENV) develops WQGs for priority substances to help with decision-making that relates to water quality. To ensure that the WQGs are transparent and scientifically defensible, ENV follows a WQG derivation protocol that establishes the type of data that are included and the process for deriving the final WQG value. Other jurisdictions that develop WQG also have derivation protocols, which vary from one another in various ways.

B.C. has updated its WQG derivation protocol (ENV 2019) to make it more consistent with the approach taken by other jurisdictions including the Canadian Council of Ministers of the Environment (CCME). WQG development is a resource intensive process and by making these changes, B.C. will have enhanced opportunities to collaborate and share guidelines with other jurisdictions.

A major feature of this new protocol is the use of species sensitivity distributions (SSDs) to estimate WQGs in addition to the deterministic method that B.C. has employed over the past decades (ENV 2012). With SSDs, the distribution of species sensitivities is modeled so that the lower 5th percentile of the distribution can be estimated and used as a basis for a guideline value (Posthuma et al. 2002). This value is referred to as the hazard concentration for five percent of the species or HC₅. Many other jurisdiction (e.g. Canada, European Union, Australia New Zealand) use SSDs to calculate WQGs (Zajdlik 2016). Although there are a number of limitations of the SSD method (ENV 2019) it is the current standard for deriving WQGs.

The CCME protocol for water quality derivation employs the use of an SSD approach when adequate data are available (CCME 2007). Along with the derivation protocol, CCME commissioned the creation of a specialized Microsoft Excel package, called SSD Master, to select the best fitting distribution to the species endpoints and compute an HC₅ to use as the guideline value (Intrinsik 2013). SSD Master uses a non-linear least squares regression (NLLSR) approach to fit each distribution (model). Although this was a reasonable approach at the time, the advent of free statistical software packages such as R (R Core Team 2017) has increased the options for fitting SSD models.

The ENV is interested in developing new methods and software to fit SSD models in support of its updated WQG derivation protocol. The purpose of this report, therefore, is to identify and propose new statistical methods for computing a HC₅ and its associated uncertainty. Specifically, this report will discuss the limitations of using a NLLSR approach to fit SSDs; investigate new approaches for calculating SSDs and estimating uncertainty; compare guideline values computed using NLLSR with those computed using a proposed new approach; and investigate statistical methods for including data point uncertainty such as within-species variability and measurement error.

There are a number of concerns about using a NLLSR approach that is used by the SSD Master Excel program: 1) it depends on arbitrary choices in computing the cumulative distribution function; 2) it fails to recognize the non-independence of the ordered end points when fitting the distribution and estimating the uncertainty of HC₅; and 3) there is little flexibility to consider measurement uncertainty or intra-species variability when there are multiple endpoints measured for a species.

Maximum likelihood estimation (MLE) is a method for fitting a statistical distribution to a univariate dataset which, as the name implies, maximizes the likelihood that the observations came from the fitted distribution. It is a very general procedure so it can provide the necessary flexibility to fit a distribution to species sensitivity data and estimate the uncertainty of the HC₅. MLE is designed for use with a univariate data set and so makes no arbitrary assumptions when fitting a distribution and it is easily

implemented using modern statistical software (such as R). MLE is used by Australia & New Zealand and EU Nations to fit the SSD and compute WQGs. Australia & New Zealand use the BurrliOz software and the EU Nations use MOSAIC_SSD, an online application (Delignette-Muller and Dutang 2015).

Although MLE has a number of advantages over the NLLSR approach, there are some issues that need to be considered when fitting SSDs with MLE. The first is the fact that multiple distributions can fit the species sensitivity data equally or almost equally well making selecting a best fitting distribution problematic. The choice of the distribution will ultimately impact the HC₅ estimate so it is important that this process is robust. A potential solution is to fit multiple distributions to the data and calculate a weighted average of the HC₅ estimate. This model averaging approach will avoid the need to select one distribution to calculate the HC₅ estimate

A comparison of HC₅ estimates calculated for seven chemicals showed that there was no consistent pattern to the HC₅ estimates calculated using the NLLSR approach compared to the model averaging MLE technique; however, there was a large difference between the uncertainty of the HC₅ estimates using the two methods. Additionally, the model averaging approach showed greater stability than single distributions when data points from the middle or right side of the distribution were added.

The second issue is the lack of methods for incorporating the uncertainty associated with the species sensitivity data. This was explored using Bayesian methods with data-cloning (Lele et al., 2007). A comparison of MLE and Bayesian methods with data-cloning lead to distributions with a similar value for the mean, but by allowing for uncertainty by using data cloning, the fit has a slightly smaller estimated standard deviation. Consequently, the estimated HC₅ from using data cloning is slightly higher than that from the fit using the geometric mean of the individual data values and the 95% confidence interval for the HC₅ is narrower when using data cloning.

The evidence supplied in this report supports the use of MLE and model averaging when deriving WQGs. These tools represent an advance in the statistical approach to deriving WQGs. Specifically, this report supports the following four conclusions:

1. MLE provides many advantages over NLLSR, the greatest of which are the accurate estimate of uncertainty associated with the HC₅ estimate and a greater flexibility.
2. Model averaging can be used to retain information obtained from multiple distributions. This is useful when more than one distribution fits the species sensitivity data set equally, or nearly equally well. It can also reduce the uncertainty associated to fitting distributions to small data sets.
3. Confidence intervals can be narrowed by using Bayesian methods with data cloning to include variability in endpoint measurements for each species.
4. Improving the statistical methods for calculating HC₅ does little to reduce the lack of theoretical support for using a HC₅ as a WQG. Care must be taken in its interpretation as it may not represent the level that is protective of 5% of species in an ecosystem or region.

CONTENTS

1. INTRODUCTION.....	1
1.1 R Code Availability	1
2. BACKGROUND: STATISTICAL METHODS FOR FITTING SPECIES SENSITIVITY DISTRIBUTIONS (SSD) AND ESTIMATING UNCERTAINTY.....	2
2.1 Description of the SSD and Associated Uncertainties	2
2.2 Limitations of the Non-Linear Least Squares Method	3
2.3 MLE Approach to Fitting an SSD	4
2.3.1 Using Tolerance Limits to Estimate Uncertainty in HC ₅	4
3. NEW APPROACHES TO CALCULATING WATER QUALITY GUIDELINES.....	5
3.1 Combining the Fit from Several Distributions using a Model Averaging Approach	5
3.2 Accounting for Uncertainty in the Species Endpoints.	6
4. METHODS.....	7
4.1 Demonstrating the Model Averaging Method	7
4.2 Comparison of Water Quality Guidelines Calculated using Two Statistical Approaches	8
4.3 Comparison of the Stability of Model Averaging with Single Distributions	8
4.4 Including Multiple Endpoints for the Same Species	8
5. RESULTS	10
5.1 Illustration of the Model Averaging Methodology	10
5.1.1 Boron	10
5.1.2 Silver	12
5.2 Comparison of MLE and Non-Linear Least Squares Regression	12
5.3 Comparison of the Stability of the Model Averaging with Single Distributions	14
5.4 Including Endpoint Uncertainty	16
6. DISCUSSION	17
6.1 Model Averaging Approach	17
6.2 Including Endpoint Uncertainty	19
6.3 Application of MLE to SSD.....	20
7. CONCLUSIONS.....	21
8. REFERENCES.....	22

ACRONYMS

AIC	Akaike information criterion
AICc	Corrected Akaike information criterion
AD	Anderson-Darling goodness of fit statistic
B.C.	British Columbia
CCME	Canadian Council of Ministers of the Environment
CFD	Cumulative frequency distribution
CRAN	Comprehensive R Archive Network
CVM	Cramer-von Mises goodness of fit statistic
ENV	B.C. Ministry of Environment and Climate Change Strategy
HC _x	Hazard concentration for x% of species
KS	Kolmogorov-Smirnov goodness of fit statistic
MCMC	Markov chain Monte Carlo
MLE	Maximum likelihood estimation
NLLSR	Non-linear least squares regression
SSD	Species sensitivity distribution
WQG	Water quality guideline

1. INTRODUCTION

Water quality guidelines (WQGs) are used as benchmarks to interpret environmental risks related to changes in water quality. In B.C., the Ministry of Environment and Climate Change Strategy (ENV) develops WQGs for priority substances to help with decision-making that relates to water quality. To ensure that the WQGs are transparent and scientifically defensible, ENV follows a WQG derivation protocol that establishes the type of data that are included and the process for deriving the final WQG value. Other jurisdictions that develop WQG also have derivation protocols, which vary from one another in various ways.

B.C. has updated its WQG derivation protocol (ENV 2019). A major feature of this new protocol is the use of species sensitivity distributions (SSDs) to estimate WQGs in addition to the deterministic method that B.C. has employed over the past decades (ENV 2012). With SSDs, the distribution of species sensitivities is modeled so that the lower 5th percentile of the distribution can be estimated and used as a basis for a guideline value (Posthuma et al. 2002). This value is referred to as the hazard concentration for five percent of the species or HC₅. Many other jurisdiction (e.g. Canada, European Union, Australia New Zealand) use SSDs to calculate WQGs (Zajdlik 2016).

The Canadian Council of Ministers of the Environment (CCME) has a protocol for water quality derivation that employs the use of an SSD approach when adequate data are available (CCME 2007). Along with the derivation protocol, CCME commissioned the creation of a specialized Microsoft Excel package, called SSD Master, to select the best fitting distribution to the species endpoints and compute an HC₅ to use as the guideline value (Intrinsik 2013). SSD Master uses a non-linear least squares regression (NLLSR) approach to fit each distribution (model). Although this was a reasonable approach at the time, the advent of free statistical software packages such as R (R Core Team 2017) has increased the options for fitting SSD models.

The ENV is interested in developing new methods and software to fit SSD models in support of its updated WQG derivation protocol. The purpose of this report, therefore, is to identify and propose new statistical methods for computing HC₅ and its associated uncertainty. Specifically, this report will discuss the limitations of using a NLLSR approach to fit SSDs; investigate new approaches for calculating SSDs and estimating uncertainty; compare guideline values computed using NLLSR with those computed using a proposed new approach; and investigate statistical methods for including data point uncertainty such as within species variability and measurement error.

1.1 R Code Availability

The package `ssdtools` was developed as a result of earlier drafts of this document and is available on the Comprehensive R Archive Network (CRAN) (Thorley and Schwarz 2018) and a shiny app (Dalgarno 2018) is available at: <https://bcgov-env.shinyapps.io/ssdtools/>.

2. BACKGROUND: STATISTICAL METHODS FOR FITTING SPECIES SENSITIVITY DISTRIBUTIONS (SSD) AND ESTIMATING UNCERTAINTY

2.1 Description of the SSD and Associated Uncertainties

An SSD is a statistical approach that fits a distribution to species sensitivity data to estimate a low percentile of the distribution (typically the lower 5% of the distribution, denoted as HC₅) which can be used as the basis for a WQG. There are a number of steps that need to be completed when taking an SSD approach to developing WQGs. These are: selecting toxicological data; fitting multiple statistical distributions to the data set; evaluating distribution fit; selecting the best fitting distribution; using the fitted distribution to estimate the 5th percentile; and estimating uncertainty of the estimate. Each of these steps are discussed further below.

Minimum data requirements to conduct an SSD are typically predetermined by each jurisdiction. Toxicological endpoints are gleaned from the literature and collated in a database. There are often multiple endpoints for each species, so the most sensitive endpoint is typically chosen and if more than one of these endpoints exists, then a geometric mean is taken to represent this species in the SSD (CCME 2007). In this way, the within-species sensitivity is not considered but rather collapsed into a single value calculated with the geometric mean.

Once the data have been selected, there are a number of statistical approaches that can be used to fit a curve to the data: NLLSR, MLE, Bayesian methods and non-parametric methods. The most commonly used methods are NLLSR and MLE so these are the focus of the discussion below.

When using NLLSR, the selected data are ordered from least to greatest and a cumulative frequency distribution (CFD) is calculated to create a second variable (cumulative percentage). This step is not necessary when using an MLE approach because MLE is designed to fit a distribution to a univariate dataset while the NLLSR requires a bivariate data set.

There is no underpinning ecological theory to determine which statistical distributions should be selected but using common sense (e.g. no negative concentrations) and patterns in the data, some distributions can be selected as good candidates (Zajdlik & Associates 2006). Starting with this list of statistical distributions, each distribution is fit to the data and the parameters of the fitted distribution are estimated. Goodness of fit statistics and diagnostic graphics are then used to select the best fitting distribution. In some situations, more than one model may fit the data equally well creating a situation with no clear answer. Also, addition or deletion of one data point may change the fit of the models. Once the model has been selected, the 5th percentile is used as the WQG. This is called the hazard concentration and is denoted as HC₅. An estimate of the uncertainty of HC₅ is then calculated.

The interpretation of the HC₅ is somewhat problematic. Despite it appearing to provide protection of a proportion of species in ecosystems (Posthuma et al. 2002), such an interpretation is not valid because the set of species used in the fitting process cannot be considered to be a random sample from the population of all species (Smith and Cairns 1993). The use of a random sample from the entire population of samples is the basis for making inference using parametric statistics. Rather than a random sample, species are selected for the SSD based on quotas (e.g. at least 5 fish species) or biological knowledge of sensitivity (certain species are included that are known to be sensitive to a chemical) and therefore only represent a selection of species that are used in laboratory tests.

2.2 Limitations of the Non-Linear Least Squares Method

In Canada, SSDs have been calculated using the SSD Master Software developed for CCME (Intrinsik 2013). SSD Master fits a number of candidate distributions to the empirical cumulative distribution function using a NLLSR method and chooses the best fit distribution to make inference on the HC₅. There are a number of concerns about using a NLLSR approach: 1) it depends on arbitrary choices in computing the cumulative distribution function; 2) it fails to recognize the non-independence of the ordered end points when fitting the distribution and estimating the uncertainty of HC₅; and 3) there is little flexibility to consider measurement uncertainty or within-species variability when there are multiple endpoints measured for a species.

The NLLSR approach determines the distribution that best fits a cumulative density plot of the species sensitivity data using NLLSR based on the distance between the fitted cumulative distribution and the observed cumulative distribution, i.e. it finds the parameters of the distribution that minimize

$$\arg \min \sum \left(\hat{F}(x_i) - F_{distr}(x_i) \right)^2$$

where $\hat{F}(x_i)$ is the empirical cumulative distribution function (i.e. based on the actual data) and $F_{distr}(x_i)$ is the theoretical cumulative distribution function.

As mentioned above, there are a number of disadvantages to the NLLSR method. First, it relies on an arbitrary choice for calculating the empirical cumulative distribution. All methods start by sorting the observed data from smallest to largest. Then the empirical cumulative probability is determined as

$$\hat{F}(x_i) = \frac{i - a}{n + 1 - a - b}$$

where $\hat{F}(x_i)$ is the empirical cumulative probability for the i^{th} observation (after sorting), n is the total sample size, and a and b are constants. There are at least six different choices for a and b , each which gives a different empirical cumulative distribution function and can therefore affects the selection of the best fit model (Bock 2015). Each choice of a and b is optimized for a particular distribution, although by convention the Hanzen method ($a = 0.5$; $b = 0.5$) or the Filiben ($a = 0.3175$; $b = 0.3175$) are most often used. This choice may affect the fit of the distributions and therefore the final HC₅.

Second, the assumptions associated with running a NLLSR test are not met when using a cumulative distribution. A key assumption of NLLSR is that each observation is independent of every other observation (Bates and Watts, 2007). This is untrue when data are sorted and an empirical cumulative distribution is used. Given this non-independence, it is unclear how to estimate the standard error (a measure of how close the estimate could be from the true unknown value of the parameter). The NLLSR approach used by the SSD Master Excel software (Intrinsik 2013) returns a 95% fiduciary limit which is obtained by inverting the confidence intervals from a non-linear least squares fit (Intrinsik 2013). There is no theoretical basis for estimating the uncertainty of the HC₅ value in this the way (Intrinsik 2013) and it is likely that the confidence intervals are too narrow given that observations are not independent when combined into an observed cumulative frequency distribution.

Third is the lack of flexibility associated with applying the NLLSR. For example, NLLSR methods based on the empirical cumulative distribution cannot easily deal with multiple endpoints for a single species. Generally, a geometric mean is used to represent these data but in doing so, information may be lost regarding the variation in within-specific sensitivity. Also, the NLLSR approach cannot include the uncertainty associated with endpoint estimates for individual species.

2.3 MLE Approach to Fitting an SSD

MLE is a method for fitting a statistical distribution to a univariate dataset which, as the name implies, maximizes the likelihood that the observations came from the fitted distribution. It is defined as:

$$L(\theta) = \prod_{i=1}^n f(x_i|\theta)$$

where x_i is the n^{th} observation of variable X and f is the density function of the parametric distribution.

As described by King et al. (2014), “The likelihood function gives the probability of observing the data given the parameters. Maximizing the likelihood implies selecting the parameters for which the probability of observing the data is highest. Maximum likelihood is by far the most standard approach to distribution fitting and more generally to model fitting.”

MLE is a very general procedure so it can provide the necessary flexibility to fit a distribution to species sensitivity data and estimate the uncertainty of the HC_5 . MLE is designed for use with a univariate data set and so makes no arbitrary assumptions when fitting a distribution; it relies on neither plotting position nor the CDF directly, but fits the distribution using the likelihood function. MLE is used by Australia & New Zealand and the European Commission to fit the SSD and compute the HC_5 . Australia & New Zealand use the BurrliOz software and the European Commission uses MOSAIC_SSD, an online application (Delignette-Muller and Dutang 2015).

2.3.1 Using Tolerance Limits to Estimate Uncertainty in HC_5

A major advantage of using MLE to generate HC_5 values is the ability to estimate the uncertainty associated with the calculated HC_5 value. Using MLE, an estimate of the uncertainty can be generated using non-parametric bootstrapping techniques. The original data are sampled with replacement, a distribution is fit to the bootstrap sample, and the fitted distribution is used to estimate the 5th percentile. This is done a large number of times (typically 1,000 times) and results in a distribution of the HC_5 values based on random resamples of the species sensitivity data. When using 95% tolerance intervals it is possible to make statements in the form “95% confident that the true HC_5 value for this distribution is within the tolerance interval” or using a one-sided lower tolerance interval ‘ x ’, then it can be stated, “95% confident that no more than 5% of the distribution fall below x ”.

3. NEW APPROACHES TO CALCULATING WATER QUALITY GUIDELINES

As described in the previous section, MLE has a number of advantages over the NLLSR approach but there are a couple of issues that need to be considered when fitting SSDs with MLE. The first is the fact that multiple distributions can fit the species sensitivity data equally or almost equally well, making the selection of a best fitting distribution problematic. The choice of the distribution will ultimately impact the HC₅ estimate so it is important that this process be robust. The second issue is the lack of methods for incorporating the uncertainty associated with the species sensitivity data. Both issues are discussed further below.

3.1 Combining the Fit from Several Distributions using a Model Averaging Approach

Often, more than one distribution will provide an adequate fit to the observed data but may have substantially different estimates of the HC₅. In a simplistic application of the NLLSR and MLE approaches, a single best-fitting distribution is chosen and used to derive the HC₅. Choosing a single “best” fit distribution can introduce a user bias if multiple distributions fit the data almost equally as well. Also, often a small change in the data will result in a different distribution being the “best” fit. By choosing the single best fitting distribution, all information in the estimated HC₅ among different distributions has been ignored.

Information theoretic approaches (Burnham and Anderson, 2002) can be used to combine the estimated HC₅ values from several distributions. This is done in the following way:

- Create a potential model set, which is the set of possible distributions to be considered (e.g. lognormal, Weibull, etc.)
- For each distribution in a model set, use maximum likelihood methods to fit the candidate model.
- For each candidate fit, compute the information theoretic value. For example, the Akaike information criterion (AIC) is a measure of the relative quality of the fit of each distribution to a given set of data. The AIC value is a trade-off between fit (measured by the maximum likelihood value) and complexity (measured by the number of parameters) and is computed as

$$AIC = 2k - 2 \log(\hat{L})$$

where k is the number of parameters of the distribution, and \hat{L} is the maximum likelihood value. A correction for small sample sizes is often applied (Burnham and Anderson, 2002) and is denoted as AIC_c .

- Rank the models using the AIC_c values. The model with the lowest AIC_c value is the best model of the model set in terms of the trade-off between fit and complexity.
- Determine the difference in AIC_c from the best fitting model of the model set, denoted as ΔAIC_c and the model weight for each distribution as

$$W_i = \frac{e^{-\Delta AIC_{ci}/2}}{\sum_{\text{All models}} e^{-\Delta AIC_{ci}/2}}$$

These model weights sum to 1 and represent the relative weight to be given to a model compared to the other models in the set. Models that do not fit the data well (relative to the other models in the set) will have large values of ΔAIC_c and a small model weight. If there are two models that equally fit the data, they will have similar ΔAIC_c and similarly model weights.

- Use the model weights to compute a weighted average of the estimated HC₅. The model weights can also be used on the tolerance intervals of the HC₅. This will automatically include the uncertainty from fitting different distributions into the derivation of the HC₅ value.

Some caution is required when selecting distributions to be fit to the data. Selected distributions must be bounded by 0 to account for the fact that the results of toxicity tests, (i.e., the effect concentrations), cannot be negative. For example, a log-Normal distribution should be used to fit the observed data rather than a Normal distribution. Conversely, when fitting the distribution on the logarithmic scale, then distributions that are bounded below by 0 should not be used, (e.g. fitting a Weibull distribution on logarithmic values).

Ideally, the observed concentrations should be used directly with the fit of the log-equivalent distribution (rather than transforming the observations using logarithms and then using the normal distribution on the logarithm of the concentration). This avoids the potential of creating a fail of fit if the units were changed. For example, the CCME guideline for boron was fit using a Gompertz distribution to the logarithm (base 10) of the concentration data. If the data were expressed in different units (e.g. µg/L rather than mg/L), this fit would then fail as some concentrations would then have values less than 1 (in the new units) and the logarithm would be negative violating the Gompertz distribution support.

3.2 Accounting for Uncertainty in the Species Endpoints.

The endpoints for species-specific toxicity tests also have uncertainty. This uncertainty can arise in a number of ways:

- Multiple endpoint values for the same species are available from a variety of studies due to either experimental error or intra-specific variation.
- The endpoint was calculated from a statistical model (e.g., a dose-response curve) and the endpoint has been estimated from this fitted curve.
- The endpoint was calculated from a mechanistic model to predict the endpoint based on endpoints from similar species using similar chemicals.

Aldenberg and Rorije (2013) and King et al. (2015) explored the impacts of ignoring the uncertainty in estimating the HC₅. Both concluded that the impact of the uncertainty in the species endpoints on the fitted curve is minor assuming that observed endpoints are centered around the true underlying endpoint. Hence the actual estimate of the HC₅ is little affected, but the uncertainty about the HC₅ is different with and without including the uncertainty of the endpoints.

The problem of including endpoint uncertainty in a direct application of MLE is the presence of latent (hidden) variables that must be integrated out. In this case, the latent variables are the true endpoint for each species. However, the true endpoint is unknown and must be estimated using the observed endpoints for each species. The following series of equations demonstrates this mathematically.

Let the true, underlying (latent) endpoints for each species (E_i) follow a distribution across species that depends upon parameters θ

$$E_i \sim f_{SSD}(E_i | \theta)$$

e.g. the distribution for the true endpoints is log-normal. However, we cannot observe the E_i directly – we only observe endpoints C_{ij} for $j=1, \dots, n_i$ that are sampled from a distribution around E_i with parameters ϕ , i.e.

$$C_{ij} \sim g_{obs}(C_{ij}|E_i, \emptyset)$$

e.g. the observed values of C could follow a normal distribution around the true (latent) endpoint. If both C_{ij} and E_i were observable, the joint distribution is simply the product of the two distributions:

$$(E_i, C_{ij}) \sim f_{SSD}(E_i|\theta) \times g_{obs}(C_{ij}|E_i, \emptyset) \quad (\text{Equation 1})$$

and the usual likelihood methods could be used as explained earlier. However, the E_i are not observable (they are latent, or hidden), and therefore this is not a proper likelihood function. Likelihood functions must depend only on observable events and cannot be a function of unobservable data. The distribution for the observable C_{ij} must integrate over the unobservable E_i

$$C_{ij} \sim \int_{E_i} f_{SSD}(E_i|\theta) \times g_{obs}(C_{ij}|E_i, \emptyset) dE_i \quad (\text{Equation 2})$$

which is computationally very difficult to implement.

Both Aldenberg and Rorije (2013) and King et al. (2015) adopted a Bayesian approach because it provides a very convenient way to evaluate the integral through Markov chain Monte Carlo (MCMC) sampling. In MCMC sampling, values from the distribution of E_i are repeat generated. For each simulated value, the joint probability distribution in Equation 1 can be evaluated. Then the integral in Equation 2 is approximated by taking the average values of Equation 1 over the generated values of E_i . The generated values of E_i will incorporate measurement error automatically.

One disadvantage of Bayesian methods is the need to specify prior distributions for the parameters of interest. The choice of prior distributions is somewhat arbitrary, and different prior distributions can lead to different estimates of the parameters of the SSD distribution. Fortunately, Lele et al. (2007) showed that a method called Data Cloning can use Bayesian methods to produce maximum likelihood estimates that are not affected by the choice of prior distribution. As the name implies, data are cloned (i.e. multiple copies are made) and run through the Bayesian analysis. The results from the Bayesian analysis are adjusted for the “artificial cloning” of the data.¹ This MLE fit can then be used in the model averaging approach described in Section 3.1.

4. METHODS

4.1 Demonstrating the Model Averaging Method

The model averaging methodology was tested using species sensitivity data extracted from the CCME Guidelines for the Protection of Aquatic Life for boron (CCME 2009) and silver (CCME 2015). SSDs were fit using MLE with the model averaging procedure described in Section 3.1. Distributions were fit using the recently developed shiny app for fitting SSD data in R² (Dalgarno 2018; Thorley and Schwarz 2018). Generally, up to five distributions were fit to each dataset: the Gompertz, log-Gumbel, log-normal, log-logistic, Gamma. These distributions were selected based on the advice of Intrinsik (2013). Cumulative distribution plots were used to visually compare the fit of each distribution. The information from

¹ Refer to <http://datacloning.org/> for an overview of the process.

² The *fitdistrplus* package for R (Delignette-Muller and Dutang 2015) was used in earlier drafts of this report but all figures and analysis have been updated using the recently released shiny app.

multiple distributions was then combined using the information theoretic methods described in Section 3.1.

4.2 Comparison of Water Quality Guidelines Calculated using Two Statistical Approaches

WQGs developed using the NLLSR approach were compared with those calculated using MLE with model averaging. WQGs calculated using the NLLSR approach were taken from seven WQGs published by the CCME since the adoption of the current protocol (CCME 2007). Species sensitivity data for these seven WQGs, the calculated HC₅ and the fiduciary limits were all taken from the CCME factsheets. The seven WQGs used were: boron (CCME 2009), cadmium (CCME 2014), chloride (CCME 2011a), endosulfan (CCME 2010), glyphosate (CCME 2012), uranium (CCME 2011b), and silver (CCME 2015). Distributions were fit to SSD data for each WQG guideline using the recently developed shiny app for the `ssdtools` package in R (Dalgarno 2018; Thorley and Schwarz 2018).

4.3 Comparison of the Stability of Model Averaging with Single Distributions

Estimated HC₅ values may be more stable to small changes in the dataset when calculated using model averaging compared with single distributions. The stability of the model averaging approach was compared with single distributions tested using three CCME datasets: silver (CCME 2015, n=9), uranium (CCME 2011b, n=13) and boron (CCME 2009, n=28). For each dataset, one random number consistent with the other data was added to the data set to simulate the addition of a new data value and the HC₅ estimate was calculated. This was done 100 times using 100 random numbers generated to represent five different sensitivity scenarios of the additional data point. The range of values for each sensitivity scenario was based on the predictions of the cumulative distribution frequency calculated using the model averaging method. The range of the five scenarios was: extremely sensitive (1-5%) sensitive (5-20%), middle (20-80%), insensitive (80-95%) and extremely insensitive (95-99%). For each scenario, random numbers within the given range were generated using the random number generator for the uniform distribution in R.

The results of the model averaging approach were compared against identical tests run on three single distributions. The top three best fitting distributions for each dataset were identified by the AICc metric. All simulations were completed using the `ssdtools` package in R (Thorley and Schwarz 2018).

4.4 Including Multiple Endpoints for the Same Species

A Bayesian analysis was done to examine whether or not including multiple data points for the same species results in a different HC₅ value and associated uncertainty compared with using the geometric mean of the data points. This analysis was conducted on the cadmium dataset developed for the B.C. cadmium WQG (Sinclair et al. 2015). The B.C. cadmium guideline was chosen for this analysis because the entire species sensitivity data set was available (i.e. multiple endpoints for each species) as opposed to the CCME guidelines which provide only the selected most sensitive endpoint or geometric mean of the multiple endpoints. Data were normalized to consistent hardness values using a linear regression presented in the cadmium guideline. For the analysis presented here, the most sensitive endpoint was selected following the hierarchy described in WQG derivation protocol (ENV (2019)). If multiple values of the most sensitive endpoint for one species were present, all values were included in the analysis to allow the inclusion of measurement error and/or intra-specific variability.

We implemented a Bayesian model similar to that of King et al. (2015) to account for measurement uncertainty. In particular, we assumed a log-normal distribution for the underlying (latent) species endpoints, i.e.,

$$E_i \sim \text{LogNormal}(\mu, \sigma) \quad \text{SSD distribution}$$

where E_i is the unknown (latent) endpoint for Species_{*i*}, and μ and σ are the mean and standard deviation (on the logarithmic scale) of the species sensitivity distribution.

The observed endpoints were assumed to also follow a log-normal distribution around the (latent) species endpoint:

$$C_{ij} \sim \text{LogNormal}(E_i, \sigma_{obs}) \quad \text{Observational process}$$

where C_{ij} are replicated endpoint measurements ($j=1, \dots, n_i$) for Species_{*i*}; and σ_{obs} is the variability in the measurement process. We implicitly assume that the measurement variability (on the logarithmic scale) is the same for all species. It is not necessary that all endpoints have replicates; the intra-species variability (on the log-scale) from species with replicated measurements is used to infer the potential intra-species variability for species with a single measurement. The model is also easily generalized where the uncertainty differs among species.

We assumed uniform distributions for the priors on the two standard deviations and a normal distribution prior for μ with a mean of 0 and a high variance as is usually done in Bayesian models (Gelman 2013).

This code was implemented in the BUGS language (Lunn et al, 2012) and run under JAGS (Plummer, 2003) in R (R Core Team 2017).

We then used data cloning (Lele et al., 2007) as implemented in the `dclone` package (Solymos, 2010) in R to obtain maximum likelihood estimates of the parameters that are not influenced by the choice of prior distributions. Values of $K=4, 8, 16,$ and 32 clones were generated, and convergence diagnostics showed that 64 clones were sufficient to reduce the impact of the prior distribution to low levels.

We compared the results to a log-normal distribution fit to the geometric mean of observed endpoints. In both cases, the HC₅ was estimated, including measures of uncertainty. We did not use a model averaging approach with the data cloning method because using the geometric mean directly leads to a conservative estimate of the uncertainty in the HC₅ which may be acceptable rather than doing the more complex data cloning fit.

5. RESULTS

5.1 Illustration of the Model Averaging Methodology

5.1.1 Boron

The data set used by CCME to calculate the boron WQG consisted of 28 species endpoints: 6 fish, 6 invertebrates, 6 amphibians, and 10 plants and algae (CCME 2009). Results of fitting the distributions (log-normal, Gompertz, Gamma, Weibull, log-logistic and log-Gumbel distributions) are presented in Table 1.

Table 1. Results of fitting several statistical distributions to the boron chronic dataset (CCME 2009). AD = Anderson-Darling; KS = Kolmogorov-Smirnov; CVM = Cramer-von Mises; AIC = Akaike Information Criterion; AICc = corrected AIC; LCL = lower confidence limit; UCL = upper confidence limit

Distribution	AD	KS	CVM	AIC	AICc	Delta AICc	Weight	HC ₅	95% LCL on HC ₅ mg L ⁻¹	95% UCL on HC ₅ mg L ⁻¹
Gompertz	0.60	0.12	0.08	237.61	238.09	0.00	0.27	1.30	0.98	2.70
Weibull	0.43	0.12	0.05	237.63	238.11	0.01	0.27	1.09	0.36	3.16
Gamma	0.44	0.12	0.06	237.63	238.11	0.02	0.27	1.08	0.30	3.60
Log-normal	0.51	0.11	0.07	239.03	239.51	1.42	0.13	1.68	0.87	3.59
Log-logistic	0.49	0.10	0.06	241.01	241.49	3.40	0.05	1.56	0.69	3.56
Log-Gumbel	0.83	0.16	0.13	244.19	244.67	6.58	0.01	1.77	1.13	3.23
Model Average								1.25	0.60	3.19

None of the distributions ranked the highest for all three goodness of fit statistics (Anderson-Darling, Kolmogorov-Smirnov, and Cramer-von Mises). The best distribution using the AIC_c criteria among the distributions in the model set was the Gompertz distribution, but the Weibull and Gamma distribution were also very competitive (ΔAIC_c virtually all the same). The HC₅ ranges from 1.08 to 1.77 mg L⁻¹ among the fitted distributions and the individual 95% confidence intervals ranged from 0.3 to 3.6 mg L⁻¹.

The model averaged HC₅ is 1.25 mg L⁻¹ and the 95% confidence intervals are 0.60 to 3.19 mg L⁻¹. A plot of fitted distributions is shown in Figure 1. Figure 2 shows the averaged distribution.

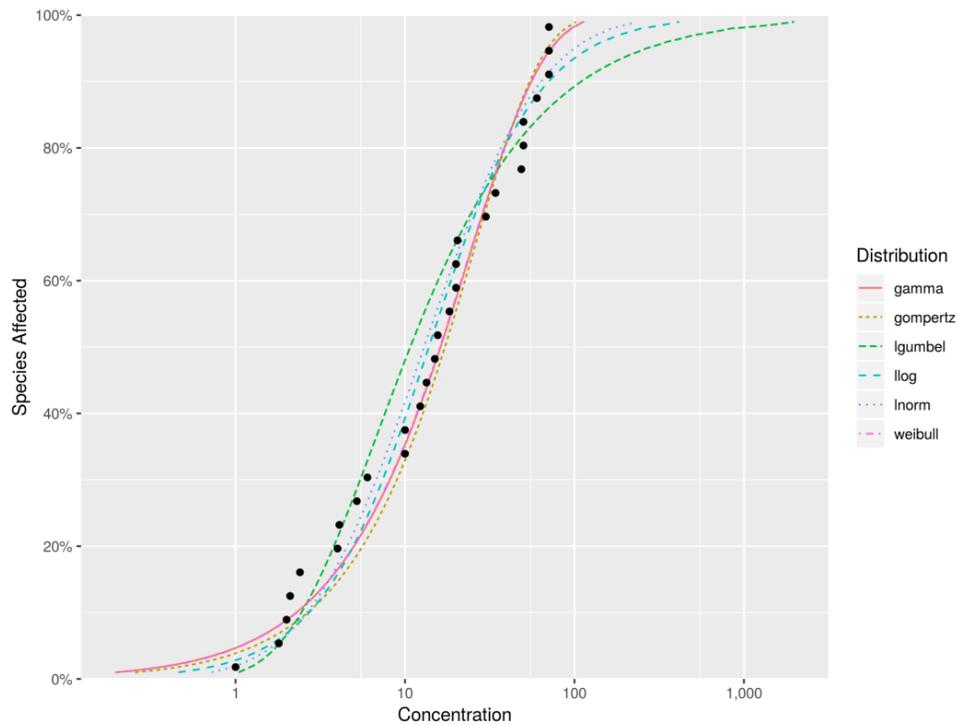


Figure 1. Comparison of several fitted distributions to the species sensitivity data for boron (CCME 2009). Data on x axis is in $\text{mg}\cdot\text{L}^{-1}$.

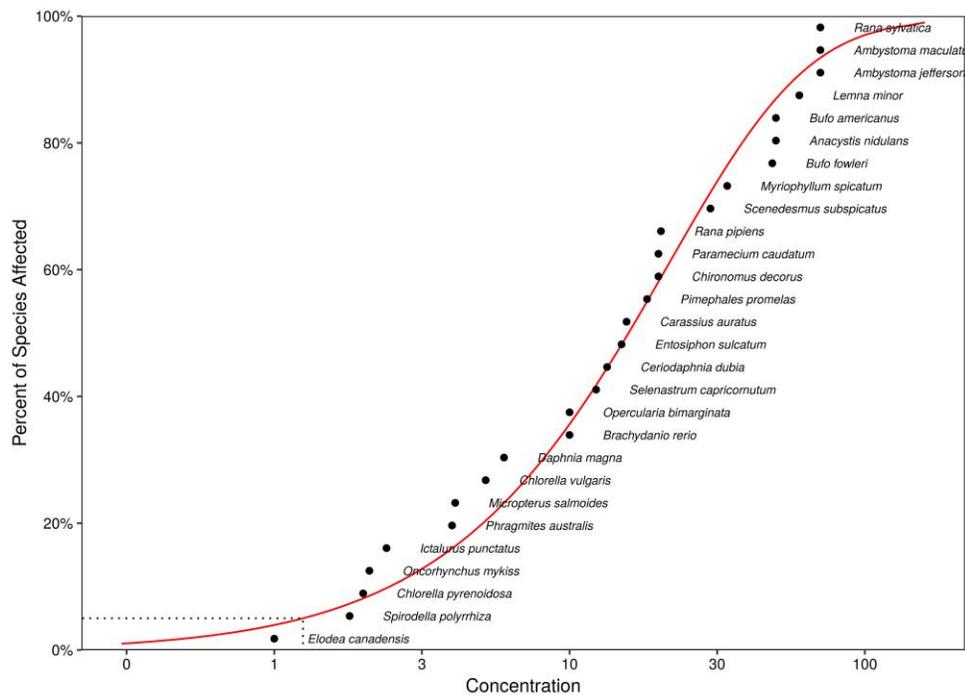


Figure 2. The weighted averaged distribution fitted to the species sensitivity data for boron (CCME 2009). Data on x axis is in $\text{mg}\cdot\text{L}^{-1}$. Averaged distributions are the Gompertz, Weibull, Gamma, Log-normal, Log-logistic and Log-Gumbel.

5.1.2 Silver

A second data set was tested using long-term CWQG extracted from the CCME Guidelines for the Protection of Aquatic Life against for silver (CCME 2015). The data consists of 9 species endpoints: 4 fish, 4 invertebrates and 1 plant. Five distributions (log-normal, log-logistic, log-Gumbel, Gamma, and Weibull) were fit to the dataset (Table 2). Both the goodness of fit statistics (Table 2) and the plot of the distributions (Figure 3) show that the log Gumbel distribution was the best fitting distribution. The model averaged distribution is given in Figure 4.

Table 2. Results of fitting several statistical distributions to the silver chronic dataset (CCME 2015). AD = Anderson-Darling; KS = Kolmogorov-Smirnov; CVM = Cramer-von Mises; AIC = Akaike Information Criterion; AICc = corrected AIC; LCL = lower confidence limit; UCL = upper confidence limit

Distribution	AD	KS	CVM	AIC	AICc	Delta AICc	Weight	HC ₅	95% LCL on HC ₅ mg L ⁻¹	95% UCL on HC ₅ mg L ⁻¹
Log-Gumbel	0.20	0.14	0.03	47.40	49.40	0.00	0.33	0.28	0.14	0.87
Log-normal	0.28	0.18	0.04	47.81	49.81	0.41	0.27	0.20	0.06	0.92
Log-logistic	0.26	0.17	0.04	48.35	50.35	0.95	0.20	0.16	0.03	0.83
Weibull	0.42	0.20	0.07	49.53	51.53	2.14	0.11	0.07	0.01	0.82
Gamma	0.55	0.24	0.10	50.12	52.12	2.72	0.08	0.06	0.00	1.09
Model Average								0.19	0.07	0.88

5.2 Comparison of MLE and Non-Linear Least Squares Regression

A comparison of guideline values calculated for seven chemicals showed that there was no consistent pattern to the guideline values calculated using the NLLSR approach compared to the model averaging MLE technique (Table 3). The average percent difference between guideline values calculated by the two methods was 37% and ranged from less than 1% for uranium to over 100% for endosulfan. There was a large difference between the uncertainty of the HC₅ estimates using the two methods. The width of the confidence interval using MLE was always greater than the fiduciary limits generated using NLLSR and ranged from 1.5 times greater for cadmium to 50 times greater for endosulfan (Figure 5).

Table 3. Results of calculating the HC₅ using the non-linear least squares regression (NLLSR) approach and model averaging using the maximum likelihood estimate (MLE). All data are from the CCME factsheets for each chemical.

Chemical	n	HC ₅ calculated with NLLSR (95% Fiduciary Limits)	HC ₅ calculated with MLE (95% CI)
Boron (mg/L)	28	1.5 (1.2-1.7)	1.2 (0.59-3.2)
Cadmium (µg/L)	36	0.09 (0.04-0.24)	0.14 (0.06-0.34)
Chloride (mg/L)	28	120 (90-150)	73 (27-198)
Endosulfan (µg ai*/L)	12	0.003 (0.0007-0.01)	0.010 (0.0012-0.51)
Glyphosate (µg ai/L)	18	800 (490-1320)	900 (459-2300)
Silver (µg/L)	9	0.25 (0.17-0.39)	0.19 (0.069-0.89)
Uranium (µg/L)	13	15 (8.5-25)	15 (3.1-120)

*ai= active ingredient

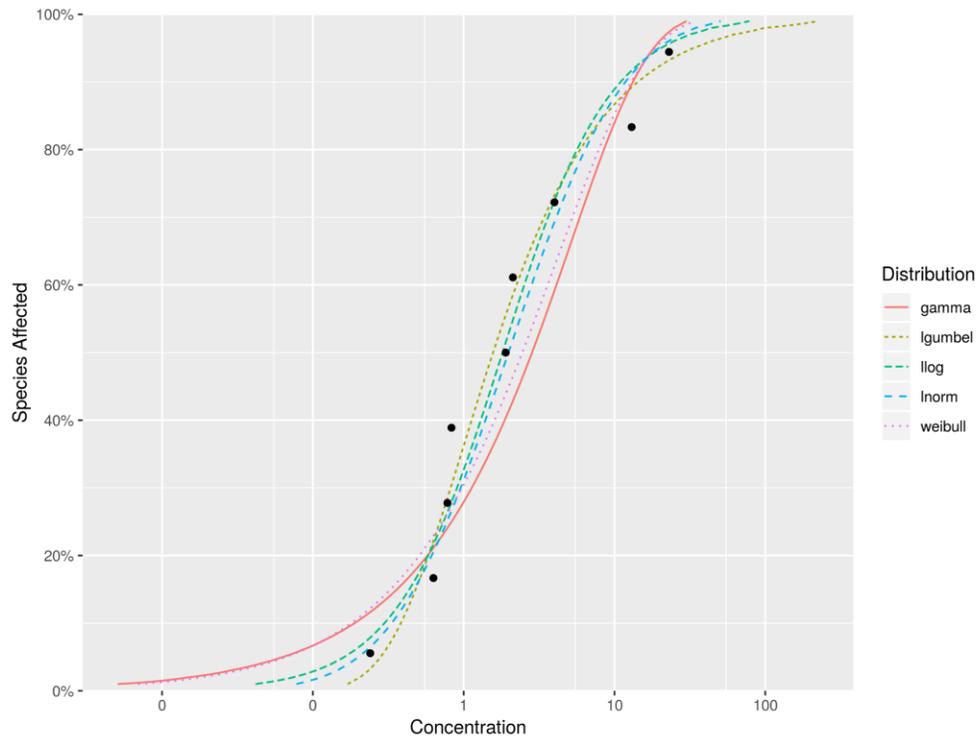


Figure 3. Results of fitting five distributions to the silver dataset (CCME 2015). There is considerable uncertainty about the HC_5 given the small number of data values.

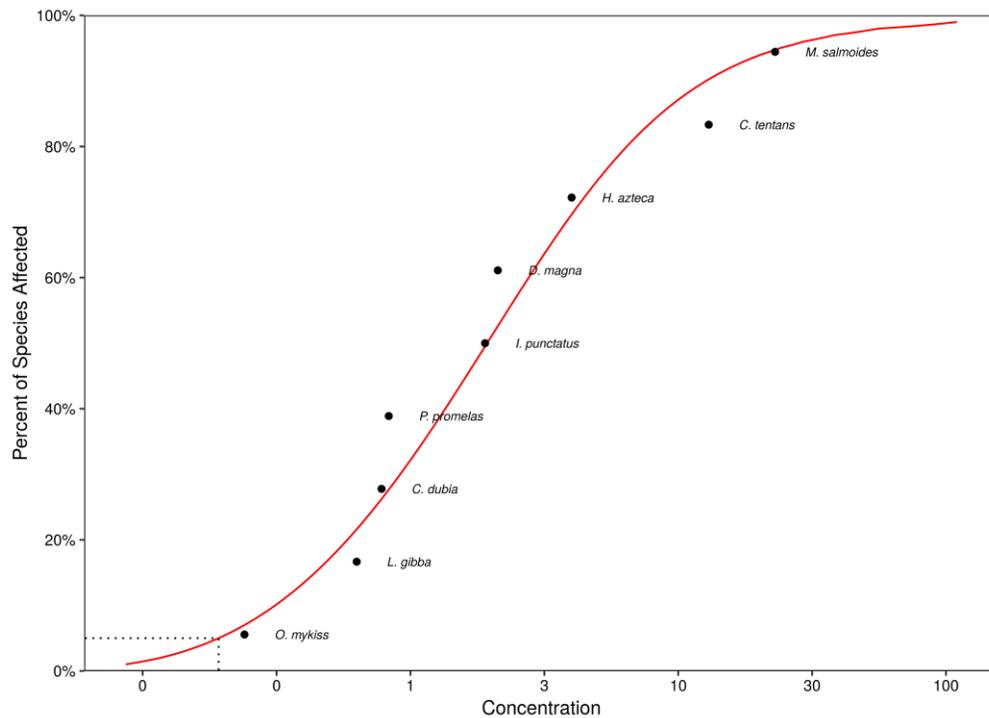


Figure 4. The weighted averaged distribution fitted to the species sensitivity data for boron (CCME 2009). Data on x axis is in $mg \cdot L^{-1}$. Averaged distributions are the Weibull, Gamma, Log-normal, Log-logistic and Log-Gumbel.

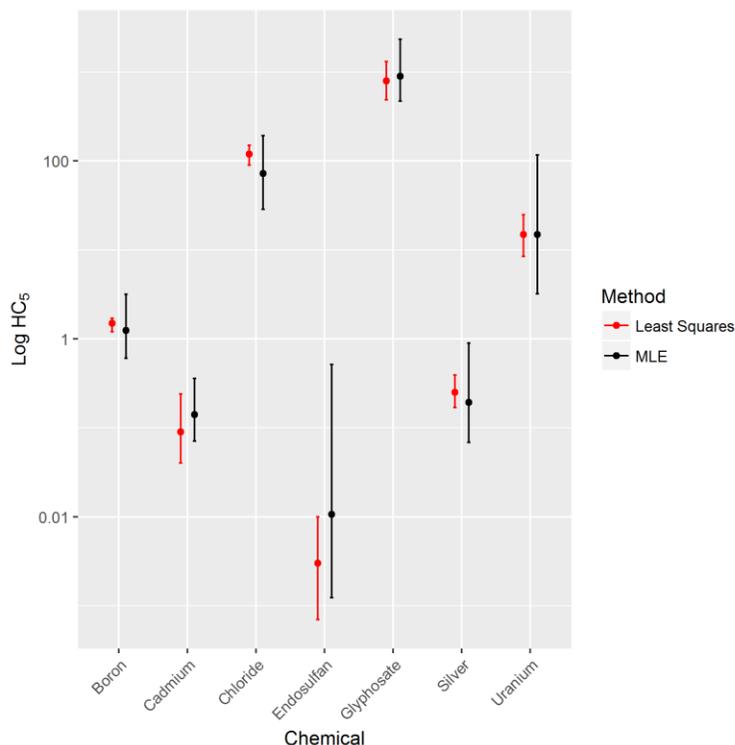


Figure 5. A comparison of HC_5 values with associated fiducial limits for non-linear least squares regression approach (NLLSR) and tolerance limits for maximum likelihood estimate approach (MLE). Note \log_{10} scale on Y axis. Units for each chemical are listed in Table 3.

5.3 Comparison of the Stability of the Model Averaging with Single Distributions

The addition of one extremely sensitive or sensitive data point decreased the median estimate of HC_5 when calculated using model averaging and the other best fitting distributions for each of the datasets tested (Figure 6). The percent decrease was as high as 49% for an extremely sensitive data point fit using the Weibull distribution to the uranium data set (Figure 6b) to 5% for a sensitive data point fit using the model averaging method with the silver data set (Figure 6c). In general, there was less of a decrease in median HC_5 estimates for the model averaging approach compared to the other distributions when a sensitive data point was added but no difference when an extremely sensitive data point was added (Figure 6). The variation in HC_5 estimates for the model averaging approach was similar or larger to the other distributions when adding a single sensitive or extremely sensitive data point.

The addition of one data point from the middle of the data set resulted in an increase in the estimate of HC_5 values for all datasets and distributions tested. The percent increase was as high as 30% for a middle point fit using the Weibull distribution to the uranium data set (Figure 6b) to as low as 4% for a middle data point fit using the model averaging method with the boron data set (Figure 6c).

All test runs with an insensitive data point produced median HC_5 estimates that were higher or equal to estimated HC_5 value calculated from the data set without the additional data point. With the addition of an extremely sensitive data point, median HC_5 estimates for most distributions decreased. This was most extreme for the Weibull distribution fit to the uranium data set (Figure 6b). The model averaging approach returned median HC_5 estimates that were within 10% of the estimated HC_5 value for all the datasets.

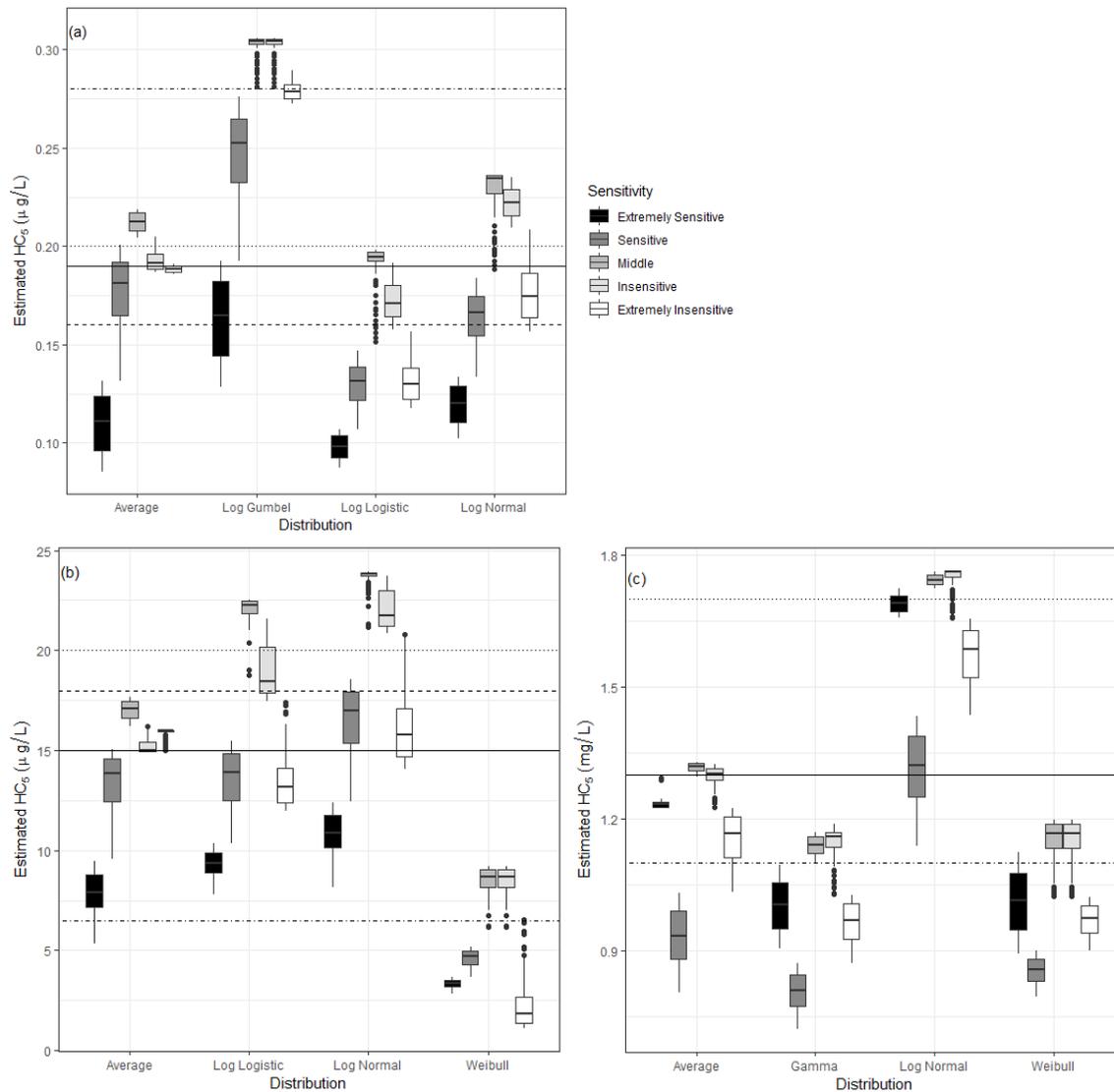


Figure 6. Results of simulations to test the stability of the model averaging approach and compare against the top three fitting distributions for a) silver (CCME 2015), b) uranium (CCME 2011b) and c) boron (CCME 2009). Horizontal lines in each figure represent the estimated HC₅ calculated by MLE for data sets without the additional data point: model averaging (solid line), log-logistic (dashed line); and log normal (dotted line). The dot-dash line is for the log Gumbel distribution in plot a, the Weibull distribution in plot b, and the gamma and Weibull distribution in plot c.

5.4 Including Endpoint Uncertainty

The observed endpoints for the effect of cadmium were extracted from Sinclair et al. (2015) (Figure 7). The estimates of the parameters and HC₅ from fitting a log-normal distribution to the geometric mean of replicate measurements and from fitting the same distribution using data cloning to the individual measurements are presented in Table 4 and shown in Figure 8.

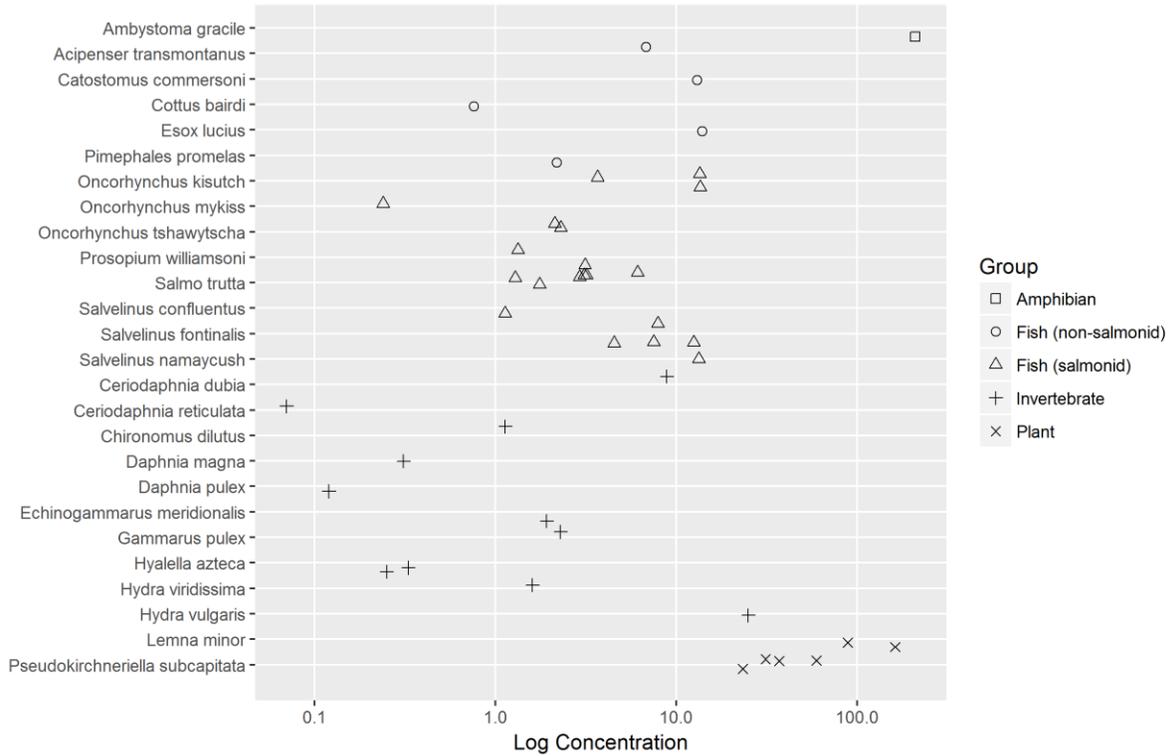


Figure 7. Low effect endpoints for species exposed to cadmium used in the MLE analysis that includes intra-specific variability. X-axis data are in log $\mu\text{g L}^{-1}$. Data are from Sinclair et al. 2015.

The two methods lead to distributions with a similar value for the mean, but by allowing for uncertainty by using data cloning, the fit has a slightly smaller estimated standard deviation (Table 4). Consequently, the estimated HC₅ from using data cloning is slightly higher than that from the fit using the geometric mean and the 95% confidence interval for the HC₅ is narrower when using data cloning. The uncertainty in each of the species endpoint allows the estimate of the underlying endpoint to be pulled slightly closer to the line of fit as seen in Figure 8.

Table 4. Comparison of estimated parameters of fitted log-normal distribution for the SSD and the estimated HC₅ from the fit using the geometric mean of the repeated measurements vs. all of the individual values. SE = standard error; SD = standard deviation; CI = confidence intervals.

	Mean (SE) $\mu\text{g L}^{-1}$	SD (SE) $\mu\text{g L}^{-1}$	HC ₅ (95% CI) $\mu\text{g L}^{-1}$
Using geometric mean of replicated data	1.14 (0.38)	1.95 (0.27)	0.13 (0.04 - 0.44)
Using data cloning and the individual measurements	1.15 (0.39)	1.91 (0.29)	0.14 (0.09 - 0.20)

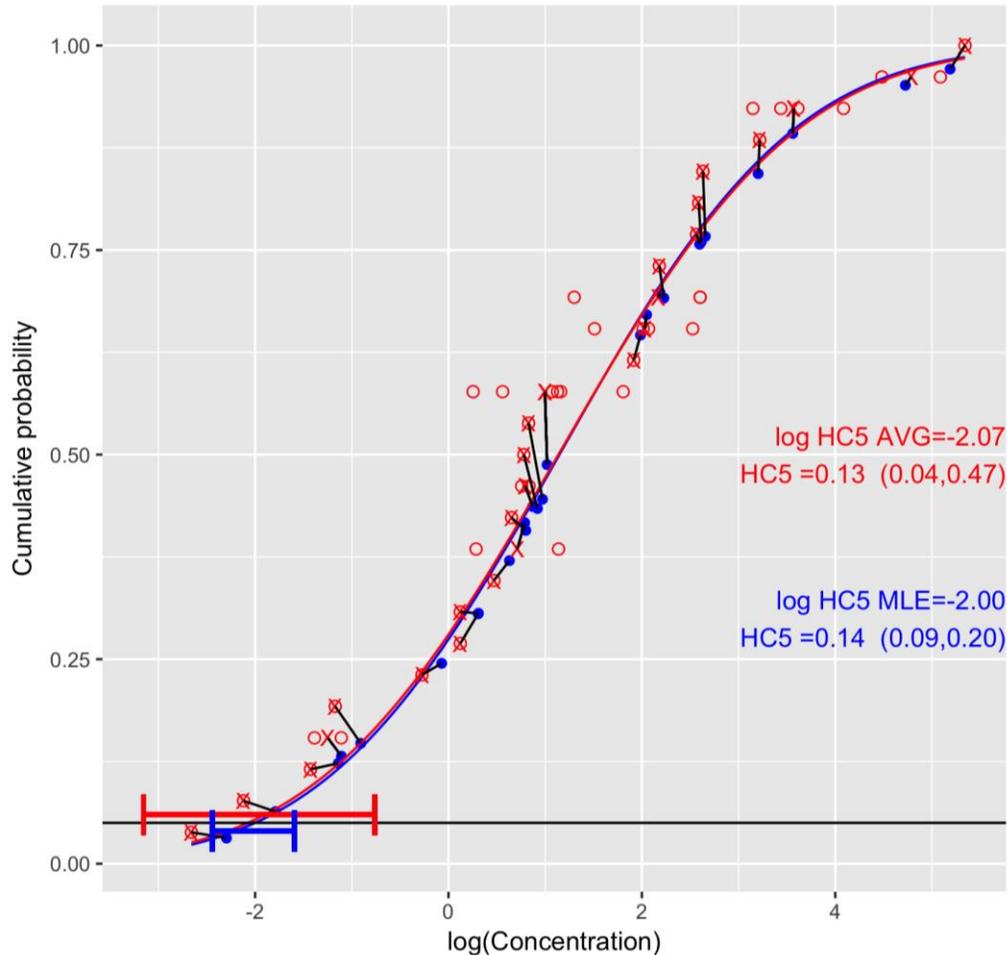


Figure 8. Comparison of the fitted log-normal SSD distribution for cadmium using the geometric mean of replicate observations (red) or all values for each species (blue). The two confidence intervals for HC_5 are shown using the same colours. The open red circles are the observed species endpoints with replicate values represented horizontally and the crosses are the geometric mean. The blue dots are the estimated underlying true endpoints (E_i) from the data cloning model that incorporate measurement error. They will generally be pulled closer to the fitted line compared to the geometric means (X 's) once measurement uncertainty is incorporated, indicated by the joining black segment.

6. DISCUSSION

6.1 Model Averaging Approach

Model averaging allows the user to fit multiple distributions to the SSD and calculate a weighted average HC_5 estimate and confidence limits. The choice of a statistical distribution can be difficult given that there is no theoretical reason to support the choice of one distribution over another. This is especially evident when a number of curves fit almost as well or when one curve fits well in the body of the data and another in the tails. The use of model averaging allows the retention and averaging of the estimates from multiple distributions, removing the need to select a single distribution.

In the case of boron, the Gompertz, Weibull and Gamma distributions all had near equal AICc values (Table 1). The selection of the log-normal distribution by CCME (2009) may have been an artifact of the plotting position used; however, the need to select a plotting position was eliminated by using MLE. The log-normal distribution is heavily constrained in how far the tails of the distribution can vary and it does not fit the center of the data very well. The Gompertz, Weibull, and Gamma distributions give a better fit to the bulk of the data though vary in how they behave in the tails of the distribution (Figure 1). The weighted averaged distribution (Figure 2) is a combination of all five of the distributions tested.

The species sensitivity dataset for silver (CCME 2015) had only nine data points resulting in considerable model uncertainty (the model weights in Table 2 are quite diffuse) especially in the fit of the tails resulting in considerable variation in the HC₅ estimates (Figure 3, Table 2). For silver, the log-Gumbel was the best fit distribution across all measures of goodness-of-fit, but the low sample size is a major source of uncertainty; an additional data point could alter the best fit distribution and resultant HC₅ estimation. Model averaging allows some buffering against this uncertainty.

Small datasets are a common concern with deriving WQGs. Part of the concern is that additional data may cause large changes in the HC₅ estimate and therefore the WQG. Model averaging does provide more stable estimates of HC₅ values compared with single distributions when an additional data point is added from the middle or right side (insensitive species) of the SSD distribution (Figure 6). For some distributions such as the log normal and log logistic, the addition of a very insensitive species can cause a decrease in the HC₅ value (Figure 6). Model averaging and single distributions respond similarly, however, with the addition of a data point from the left side (sensitive species) of the distribution: the HC₅ estimate is reduced.

Another major source of instability in the HC₅ estimate calculated using single distributions is that the best fitting distribution(s) may change with small changes in the dataset. This could cause a large jump in the HC₅ values given small changes in the dataset. The likelihood of different distributions having a better fit with small changes in the dataset was not tested in section 5.3, however the model averaging approach has recently been used by the CCME to prevent discontinuities when fitting multiple distributions across data normalized to different water chemistry conditions (CCME 2019).

A comparison of HC₅ values calculated for seven chemicals using MLE with model averaging and NLLSR revealed that one approach does not systematically give larger estimates compared to the other. For each chemical, however, MLE produces larger estimates of the uncertainty associated with the HC₅ values. As described in Section 2.2, the fiduciary limits produced by the NLLSR fit are calculated using a method with no statistical theoretical basis (Intrinsik 2013) and therefore they cannot be validated. In contrast, using MLE, the uncertainty of the estimate can be estimated using a standard non-parametric bootstrapping technique.

ECCC undertook a separate, comparative analysis between guidelines derived from SSD Master and HC₅ estimates from two MLE-based models, MOSAIC and Burrlioz (K. McTavish unpublished data, 2019). Twenty toxicity data sets were considered for a total of 60 comparisons. For 67% of the comparisons, SSD Master yielded more conservative HC₅ estimates than those from the MLE-based models and 88% of HC₅ estimates among SSD Master and the MLE-based models were generally within a factor of 2. This is noteworthy since it is generally accepted that successful model validation is achieved when simulated results are within a factor of two of the observed results (Smith et al. 1991, Parrish et al. 1992, Zacharias and Heatwole 1994).

MLE may have a positive bias compared to other distribution fitting methods with small sample sizes. For example, Etersson (2011) investigated, among other things, how different estimation methods (e.g. MLE, moment estimation, least squares) influence the bias and variance of the resulting estimated HC₅

values. A theoretical 'true' HC_5 was calculated using all the available toxicity data for specific organophosphorus and organocarbamate chemicals. Additional HC_5 values were subsequently estimated by taking 1000 random subsamples of the toxicity dataset of either 8 genus mean acute values (GMAVs) or three species mean acute values that met or didn't meet the minimum data requirements set by the 1985 U.S. EPA guideline protocol (US EPA 1995) or the Office of Pesticide Programs (OPP) benchmark species, respectively. Bias of the HC_5 was determined by taking the ratio of the estimated HC_5 to the theoretical (or true) HC_5 . Results from this study revealed that MLE consistently overestimated the true HC_5 . This is consistent with the results of the ECCC comparison (K. McTavish unpublished data, 2019).

Although MLE may have a positive bias compared with NLLSR with small sample sizes, there are several advantages of using MLE. First, the MLE approach does not use the cumulative distribution function directly so an arbitrary decision needs to be made regarding its calculation. When using MLE, each observation contributes equally to the likelihood estimate which summarizes all the information about the distribution contained in the data. Second, standard MLE methods provide estimates of uncertainty based on the likelihood function and so provide a true measure of uncertainty. The ad hoc approach used in the NLLSR method is based on the assumption of independence among the observations which is incorrect, given that the values are sorted to create the cumulative distribution curve. Third, MLE is a flexible tool that can be combined to conduct more complex analysis. For example, information theoretic methods can be used in conjunction with MLE to combine estimates of HC_5 from several distributions. The information theoretic methods automatically account for the relative support of the data to the different distributions through the model weights. The information theoretic approach can also account for model uncertainty when estimating the tolerance bounds for the percentile.

6.2 Including Endpoint Uncertainty

There is a degree of uncertainty associated with the toxicity test endpoints whether this stems from experimental error, intra-specific variation or the inherent error associated with calculating endpoints from dose response curves. The current practise is to ignore the uncertainty and use an average (e.g. geometric mean) directly in the model fitting procedure. While simplifying the problem, this fails to include the uncertainty in the individual endpoints. A direct likelihood evaluation is impractical because of the difficulty in directly computing the likelihood. However, there are two related methods to incorporate this uncertainty. The first is the use of Bayesian methods to directly incorporate both measures of uncertainty. Modern Bayesian methods rely on MCMC sampling to evaluate the hidden integrals. Some care is needed with Bayesian methods because of the potential influence of the prior distribution on the final estimates. The second method uses data cloning and Bayesian methods to obtain maximum likelihood estimates. The artificial duplication of the data enables estimates to be obtained that are free of the influence of the prior distribution.

In our one example using data cloning and Bayesian methods combined, the effect of uncertainty in the endpoints on the estimate of HC_5 was relatively small. The maximum likelihood estimate of the location parameter (the mean on the log-scale) is the simple arithmetic mean of the $\log(E_i)$ which are unobservable. However, the individual C_{ij} are assumed to follow a log-normal distribution around the $\log(E_i)$ and so the expected value of each $\log(C_{ij})$ is $\log(E_i)$ and the arithmetic mean of the $\log(\text{geometric means})$ is an unbiased estimator for location parameter. Hence, the two approaches have similar estimates for the location parameter. This will be true for any distribution whose MLE for the location parameter is a linear function of the $\log(E_i)$.

The maximum likelihood estimate for the scale parameter (the standard deviation on the logarithm scale) will be affected by measurement error if the measurement error standard deviation is large relative to the standard deviation of the E_i and there are few replicate values. In this case the measurement error standard deviation (on the log-scale) is 0.50, which is about $\frac{1}{4}$ of the standard deviation in the E_i of 1.9, with an average of 1.67 replicates per species. This implies that the error in the geometric mean around the true E_i is $0.50 / \sqrt{1.67} = 0.38$ which is 20% of the standard deviation and so the effect is minimal when the geometric means are used to estimate the standard deviation vs. using the standard deviation of the (latent) E_i (e.g. in Table 4 the estimated standard deviation using the geometric means is only slight larger than the MLE of the standard deviation from data cloning). Hence for this scale-location distribution (on the log-scale), the effect of measurement error is negligible on the estimates of the SSD distribution. In fact, as shown by Aldenberg and Rorije (2013), increasing measurement error will lead to smaller estimates of the SSD distributions standard deviation after correcting for measurement uncertainty. Intuitively, with measurement error, the spread in the species geometric means is a mixture of the underlying standard deviation and measurement error. If the measurement error is known, then the actual standard deviation of the E_i must be smaller.

These results are similar to those reported by Aldenberg and Rorije (2013, Table 1) where a small measurement error had negligible effect on estimates of the parameters of the SSD distribution.

By the same reasoning, we can show that the effect of measurement uncertainty in this example on the standard error of the estimates is also negligible (See Table 4).

The HC_5 is estimated as $\exp(\hat{\mu} - 1.645\hat{\sigma})$ where $\hat{\mu}$ and $\hat{\sigma}$ are the estimated mean and standard deviation (on the logarithmic scale). Because these estimates were affected to only a very small degree by measurement uncertainty in this case, the estimate of HC_5 is relatively unaffected.

So, where does the incorporation of uncertainty have an effect? The estimators of the scale and location using the geometric mean of repeat observations are not correlated, but the estimators from the MLE determined using data cloning have a negative sampling correlation. This means that the sampling variation of $\widehat{HC}_5 = \hat{\mu} - 1.645\hat{\sigma}$ will be smaller for the MLE using the individual concentrations compared to using the geometric means and so the tolerance interval will be narrower. This matches the results of Aldenberg and Rorije (2013):

“... the fact that accounting for data error yields less conservative SSD fits – it follows that by not taking data error into account,... one stays on the safe side. Intuitively, one would think that data error makes things worse, but that is not the case. Better measurement data, or better data predictions, reduce this conservativeness.”

While these results were shown for a log-normal distribution, we expect similar results for other distributions. Aldenberg and Rorije (2013) reached the conclusion that no corrections for measurement error are warranted given the small differences in the estimates of the HC_5 values. However, including the uncertainty in the endpoint estimates via data cloning produces estimates with narrower confidence intervals which may be of benefit to the application of WQGs in water quality management.

6.3 Application of MLE to SSD

Some care must be taken when choosing a candidate distribution to ensure the data falls within the range of the statistical distribution. Concentration values are non-negative and so any distribution must only have positive support, e.g. a log-normal distribution has support on the positive real line, but a normal distribution has support on both positive and negative values. Similarly, if fitting distributions

directly to the log(concentration) values, distributions must be chosen that have support over the full real line. For example, a Weibull distribution should not be fit to log(concentration) data.

Finally, even though statistical methods are used to derive the HC₅, caution is needed in its interpretation. This is largely because the set of species used in fitting the distribution are not a random sample from the population of species in the ecosystem to be managed, so it is incorrect to interpret the HC₅ as “protective of all but 5% of species” in a given ecosystem or across a country (Smith and Cairns 1993). As SSDs are based solely on single species toxicity tests, they do not account for inter-specific interactions that can alter species sensitivities and therefore the estimate of an HC₅ (Larras et al. 2015). The theoretical foundations of using SSDs to estimate hazard concentrations is extremely limited (Forbes and Calow 2002; Posthuma et al. 2002).

7. CONCLUSIONS

The evidence supplied in this report supports the use of MLE and model averaging when deriving WQGs. These tools represent an advance in the statistical approach to deriving WQGs. Specifically, this report supports the following four conclusions:

1. MLE provides many advantages over NLLSR, the greatest of which are the accurate estimate of uncertainty associated with the HC₅ estimate and a greater flexibility.
2. Model averaging can be used to retain information obtained from multiple distributions. This is useful when more than one distribution fits the species sensitivity data set equally, or nearly equally well. It can also reduce the uncertainty associated to fitting distributions to small data sets.
3. Confidence intervals can be narrowed by using Bayesian methods with data cloning to include variability in endpoint measurements for each species.
4. Improving the statistical methods for calculating HC₅ does little to reduce the lack of theoretical support for using a HC₅ as a WQG. Care must be taken in its interpretation as it may not represent the level that is protective of 5% of species in an ecosystem or region.

8. REFERENCES

- Aldenberg and Rorije (2013). Species Sensitivity Distribution Estimation from uncertain (QSAR-based) Effects Data. *Alternatives to Laboratory Animals*, 41 19-31.
- Bates, D. M. and Watts, D. G. (2007). *Nonlinear regression Analysis and its Applications*. Wiley, New York.
- Bock, M. 2015. Statistical tools to evaluate species sensitivity distributions and calculate final acute and chronic values. Presentation to the US EPA. Available at: https://www.epa.gov/sites/production/files/2016-01/documents/07_bock_ssd_v5_secure.pdf
- Burnham, K. P.; Anderson, D. R. (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.), Springer-Verlag, ISBN 0-387-95364-7.
- Cairns, J. 1986. The Myth of the Most Sensitive Species. *BioScience* 36:670–672.
- CCME (Canadian Council for Ministers of the Environment) 2007. A protocol for the Derivation of Water Quality Guidelines for the Protection of Aquatic Life 2007. In: *Canadian Environmental Quality Guidelines, 1999*, Canadian Council of Minister of the Environment, 1999, Winnipeg, Manitoba.
- CCME (Canadian Council of Ministers of the Environment). 2009. Canadian water quality guidelines for the protection of aquatic life: Boron. In: *Canadian environmental quality guidelines, 1999*. Canadian Council of the Ministers of the Environment, Winnipeg. Available at: <http://cegg-rcqe.ccme.ca/download/en/324/>
- CCME (Canadian Council of Ministers of the Environment). 2010. Canadian water quality guidelines for the protection of aquatic life: Endosulfan. In: *Canadian environmental quality guidelines, 1999*. Canadian Council of the Ministers of the Environment, Winnipeg. Available at: <http://cegg-rcqe.ccme.ca/download/en/327/>
- CCME (Canadian Council of Ministers of the Environment). 2011a. Canadian water quality guidelines for the protection of aquatic life: Chloride. In: *Canadian environmental quality guidelines, 1999*. Canadian Council of the Ministers of the Environment, Winnipeg. Available at: <http://cegg-rcqe.ccme.ca/download/en/337/>
- CCME (Canadian Council of Ministers of the Environment). 2011b. Canadian water quality guidelines for the protection of aquatic life: Uranium. In: *Canadian environmental quality guidelines, 1999*. Canadian Council of the Ministers of the Environment, Winnipeg. Available at: <http://cegg-rcqe.ccme.ca/download/en/328/>
- CCME (Canadian Council of Ministers of the Environment). 2012. Canadian water quality guidelines for the protection of aquatic life: Glyphosate. In: *Canadian environmental quality guidelines, 1999*. Canadian Council of the Ministers of the Environment, Winnipeg. Available at: <http://cegg-rcqe.ccme.ca/download/en/182/>
- CCME (Canadian Council of Ministers of the Environment). 2014. Canadian water quality guidelines for the protection of aquatic life: Cadmium. In: *Canadian environmental quality guidelines, 1999*. Canadian Council of the Ministers of the Environment, Winnipeg. Available at: <http://cegg-rcqe.ccme.ca/download/en/148/>
- CCME (Canadian Council of Ministers of the Environment). 2015. Canadian water quality guidelines for the protection of aquatic life: Silver. In: *Canadian environmental quality guidelines, 1999*. Canadian Council of the Ministers of the Environment, Winnipeg. Available at: <http://cegg-rcqe.ccme.ca/download/en/355/>
- CCME (Canadian Council of Ministers of the Environment). 2019. Canadian water quality guidelines for the protection of aquatic life: Manganese. In: *Canadian environmental quality guidelines, 1999*. Canadian Council of the Ministers of the Environment, Winnipeg. Available at:
- Dalgarno, D. 2018. *ssdtools: A shiny web app to analyse species sensitivity distributions*. Prepared by Poisson Consulting for the Ministry of the Environment, British Columbia. <https://poissonconsulting.shinyapps.io/ssdtools/>
- Delignette-Muller, M. L., C. Dutang, and others. 2015. *fitdistrplus: An R package for fitting distributions*. *Journal of Statistical Software* 64:1–34.

- ENV (BC Ministry of Environment). 2012. Derivation of water quality guidelines to protect aquatic life in British Columbia. Water Protection and Sustainability Branch. Available at: https://www2.gov.bc.ca/assets/gov/environment/air-land-water/water/waterquality/water-quality-guidelines/derivation-protocol/bc_wqg_aquatic_life_derivation_protocol.pdf
- ENV (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.
- Forbes, T. L., and V. E. Forbes. 1993. A Critique of the Use of Distribution-Based Extrapolation Models in Ecotoxicology. *Functional Ecology* 7:249–254.
- Forbes, V. E., and P. Calow. 2002. Species Sensitivity Distributions Revisited: A Critical Appraisal. *Human and Ecological Risk Assessment: An International Journal* 8:473–492.
- Gelman, A, Carlin, J.B., Stern, H.S., Dunson, D. B., Vehtari, A. and Rubin, D.R. (2013). *Bayesian Data Analysis*, 3rd Edition. Chapman and Hall/CRC.
- Intrinsik Environmental Sciences Inc. 2013. Determination of hazardous concentrations with species sensitivity distributions. SSD MASTER Version 3.0. Report Prepared for CCME. May 2013
- King, G. K. K., P. Veber, S. Charles, and M. L. Delignette-Muller. 2014. MOSAIC_SSD: A new web tool for species sensitivity distribution to include censored data by maximum likelihood. *Environmental Toxicology and Chemistry* 33:2133–2139.
- King, G.K.K., Larras, F., Gharles, S. and Delignette-Muller, M.L. 2015. Hierarchical modelling of species sensitivity distributions: development and application to the case of diatoms exposed to several herbicides. *Ecotoxicology and Environmental Safety*, 114, 212-221.
- Larras, F., V. Gregorio, A. Bouchez, B. Montuelle, and N. Chèvre. 2015. Comparison of specific versus literature species sensitivity distributions for herbicides risk assessment. *Environmental Science and Pollution Research* 23:3042–3052.
- Lele, S.R., B. Dennis and F. Lutscher, 2007. Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters* 10, 551–563.
- Lunn, D., Jackson, C., Best, N., Thomas, A. and Spiegelhalter, D. (2012). *The BUGS Book – A practical introduction to Bayesian Analysis*. Chapman and Hall/CRC Press.
- Parrish, R.S., Smith, C.N., and Fong F.K. 1992. Tests of the Pesticide Root Zone Model and the Aggregate Model for transport and transformation of aldicarb, metolachlor, and bromide. *Journal of Environmental Quality*, 21: 685-697.
- Posthuma, L., G. W. Suter, and T. P. Traas, editors. 2002. *Species sensitivity distributions in ecotoxicology*. Lewis Publishers, Boca Raton, Fla.
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, March 20–22, Vienna, Austria. ISSN 1609-395X.
- R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Sinclair, J.A., Schein, A., Wainwright, M.W., Prencipe, H.J., MacDonald, D.D., Haines, M.L, and Meays, C. 2015. Ambient water quality guidelines for cadmium – technical report. Water Protection and Sustainability Branch, BC Ministry of Environment. Available at: <https://www2.gov.bc.ca/assets/gov/environment/air-land-water/water/waterquality/water-quality-guidelines/approved-wqgs/cadmium/cadmium.pdf>
- Smith, E. P., and J. Cairns. 1993. Extrapolation methods for setting ecological standards for water quality: statistical and ecological concerns. *Ecotoxicology* 2:203–219.

- Smith MC, Bottcher AB, Campbell KL, Thomas DL. 1991. Field testing and comparison of the PRZM and GLEAMS models. *Transactions of the American Society of Agricultural Engineers*, 34: 838 – 847.
- Solymos, P. 2010. dclone: Data Cloning in R. *The R Journal* 2(2), 29-37. <http://journal.r-project.org/>.Zajdlik & Associates Inc. 2006. Potential statistical models for describing species sensitivity distributions. Canadian Council of Ministers of Environment. Available at: http://www.ccme.ca/files/Resources/supporting_scientific_documents/pn_1415_e.pdf
- Thorley, J. and Schwarz, C. 2018. ssdtools: Species Sensitivity Distributions. R package version 0.0.1.9002.
- US Environmental Protection Agency (EPA) 1985. Guidelines for deriving numerical national water quality criteria for the protection of aquatic organisms and their uses. Environmental Research Laboratories.
- Zacharias S, Heatwole CD. 1994. Evaluation of GLEAMS and PRZM for predicting pesticide leaching under field conditions. *Transactions of the American Society of Agricultural Engineers*, 37: 439 – 451.
- Zajdlik, B. A. 2016. A statistical evaluation of the safety factor and species sensitivity distribution approaches to deriving environmental quality guidelines. *Integrated Environmental Assessment and Management* 12:380–387.