Risk Analysis DOI: 10.1111/j.1539-6924.2011.01728.x

Competing Statistical Methods for the Fitting of Normal Species Sensitivity Distributions: Recommendations for Practitioners

Graeme L. Hickey and Peter S. Craig*

A species sensitivity distribution (SSD) models data on toxicity of a specific toxicant to species in a defined assemblage. SSDs are typically assumed to be parametric, despite noteworthy criticism, with a standard proposal being the log-normal distribution. Recently, and confusingly, there have emerged different statistical methods in the ecotoxicological risk assessment literature, independent of the distributional assumption, for fitting SSDs to toxicity data with the overall aim of estimating the concentration of the toxicant that is hazardous to p% of the biological assemblage (usually with p small). We analyze two such estimators derived from simple linear regression applied to the ordered log-transformed toxicity data values and probit transformed rank-based plotting positions. These are compared to the more intuitive and statistically defensible confidence limit-based estimator. We conclude based on a large-scale simulation study that the latter estimator should be used in typical assessments where a pointwise value of the hazardous concentration is required.

KEY WORDS: Ecotoxicological risk assessment; hazardous concentration; species sensitivity distribution

1. INTRODUCTION

The species sensitivity distribution (SSD[1]) concept is firmly embedded in regulatory and governmental ecotoxicological risk assessment programs.[2–11] An SSD is a probabilistic model of interspecies variability in sensitivity (measured as the *toxicity* to a species) to a particular toxicant (e.g., a pesticide or general chemical) for a biological assemblage, which, according to Posthuma *et al*.[1] (pp. 570–571), may be defined as a taxon, assemblage, or community.

SSDs are generally motivated by the need to estimate the environmental level of concern (ELC) in a particular ecosystem for a toxicant, but are also used to infer community-level effects.[12] For risk assessment of chemicals within the European Union[9] (EU), the ELC is taken to be the predicted no-effect concentration (PNEC). The precursor to probabilistically estimated ELCs was the deterministic assessment factor approach. In its most basic form, this consists of dividing the lowest observed species toxicity value, a median lethal (effect) concentration, or a no-observed effect concentration by an arbitrarily specified assessment factor. The magnitude of the assessment factor depends on the data type, composition, and sample size; see ECHA[9] (Table R.10-5) for an example set of assessment factors for use in deriving PNECs for aquatic compartments. The role of SSDs is therefore to allow risk assessors to better estimate ELCs by quantifying uncertainty.

In the context of chemical risk assessment within the EU,[9] the PNEC is determined as the ratio of: (1) the concentration that is hazardous to 5% of species

Department of Mathematical Sciences, Durham University, UK.
*Address correspondence to Peter S. Craig, Department of Mathematical Sciences, Durham University, UK; p.s.craig@durham.ac.uk.

(denoted as the HC_5) in the respective assemblage; and (2) a deterministic assessment factor between 1 and 5. The magnitude of (2) is chosen on an *ad hoc* qualitative basis by the risk manager as opposed to by a statistical criterion; we do not discuss it further. There are many suggested approaches for estimating (1), or more generally the HC_p for some choice of $0 < p < 100$. The large number of estimators arises from the following: (i) differing statistical foundational viewpoints, for example, classical confidence, or Bayesian credible, interval approaches[13] or a decision-theoretic analysis;[14] (ii) the choice of parametric distribution (which can incorporate extensions beyond the default forms) or whether nonparametric methods are used;[15] and (iii) assessment specific considerations, for example, the type of data used. Clearly (i) and (ii) are statistical considerations. Point (iii) on the other hand is predominantly a scientific matter, and a significant portion of the criticism pertaining to the ecological interpretability and risk assessment applicability of SSDs appears to stem from it. The lack of consensus regarding (ii) is clear, with advocates of parametric models favoring a variety of simple univariate distributions: the log-normal distribution;[16] the log-logistic distribution;[17] the log-triangular distribution;[2] and the Burr Type III distribution.[18] Hypothesis testing is the standard approach to evaluating the acceptability of such distributions; however, its effectiveness is often limited by the small sample sizes of species toxicity data available. Parametric flexibility notwithstanding, ECHA[9] (p. 22) describes the log-normal SSD representation as ''a pragmatic choice from the possible families of distributions because of the available description of its mathematical properties.''

Despite widespread application of SSDs in regulatory ecological risk assessment, the scientific community has criticized their use. As well as debating the statistical assumptions concerning distribution shape and choice of estimator, the relevance of inferences drawn from SSDs has been questioned[15,19,20] given the other assumptions made, which include: laboratory tested species are representative of untested species; toxicological endpoints measured are ecologically relevant; exposure is identical for all species; interactions between species are unimportant; all species have equal ecological importance; and a 95% protection goal (i.e., affecting $p = 5\%$ of species) is acceptable. This list is by no means exhaustive. Choosing 95% is effectively a policy decision;[21] it has been noted[22] that this is effectively inapplicable to communities of high conservation value since the loss of a single species may be unacceptable. An alternative proposed statistical method,[23] aimed at protecting the most sensitive species with a fixed level of confidence, was considered impractical[24] since early results led to ELCs below background levels.

An alternative to formal hypothesis testing for evaluating whether the sample of species toxicity data is (approximately) from a log-normal distribution is by inspection of a quantile–quantile (Q–Q) plot, otherwise referred to as a normal probability plot.[25] This approach has been described[26] as a "graphical method." A Q–Q plot for assessing log-normality is constructed by plotting the numerically ordered log-transformed toxicity values against the corresponding theoretical quantiles of the standard normal distribution. If the distributional hypothesis is correct, the points should lie approximately on a straight line; significant deviation of points from this line would suggest departures from log-normality. To complicate matters, the theoretical quantiles are dependent on the method of specification, which is usually via a choice of uniform plotting positions.

Recently,[12,27–30] linear regression models have been fitted to data in Q–Q plots in order to estimate SSDs and, subsequently, HC_p values based on the assumption of a log-normal SSD. Regression models on Q–Q plots are increasingly popular in environmental data analysis, especially when there are nondetects (i.e., censored data), since they provide convenient methods for estimating distribution parameters.[31] However, it is generally the case that censored data are either discarded or transformed into a pointwise value on an *ad hoc* basis before fitting an SSD.[32] Notwithstanding this popularity, we do not know of any research that analyzes the statistical properties of linear-regression-based HC_p predictions in comparison to parametric predictions currently employed in EU regulatory applications for noncensored data.

Despite noteworthy criticism referred to here, the SSD and HC_p concepts are repeatedly used in regulatory risk assessment and ecological research. Given this fact, it is important to support practitioners (i.e., risk assessors, research scientists, and stakeholders) using SSDs by making recommendations based on sound statistical reasoning and to expose any inadequacies in current practice. Therefore, we find analytic formulas for two regression-based HC_p estimators that have been proposed

for use with log-normal SSDs. A large-scale simulation exercise is performed to calculate estimator performance for a wide range of sample sizes. These two rules are then compared to the rule proposed by Aldenberg and Jaworska[13] that was derived to have well-defined fixed coverage properties both for repeated sampling in the frequentist (classical) statistical paradigm and with respect to the posterior distribution of the log mean and variance parameters in the Bayesian paradigm.

2. ESTIMATORS

Suppose that, for a given toxicant, a sample of n log-transformed (base 10) distinct species toxicity values (e.g., median effect concentration values) are observed, which we denote as $\mathbf{Y} = (y_1, y_2, \ldots, y_n)$. We will assume a log-normal SSD, that is, that y_i $(i = 1, \ldots, n)$ are independent and identically distributed observations from a normal distribution with mean μ and variance σ^2. We denote the numerically ordered log concentrations as $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$.

To construct a normal Q–Q plot, the theoretical quantiles $x_1, \ldots, x_n$ must be calculated; the plot then consists of points with coordinates $(x_i, y_{(i)})$ for $i = 1, \ldots, n$. There are many suggested approaches to determining the x_i, the most standard of which is to approximate them as $x_1 = \Phi^{-1}(p_1)$, $x_2 = \Phi^{-1}(p_2), \ldots, x_n = \Phi^{-1}(p_n)$, where $\Phi^{-1}(\cdot)$ is the standard normal quantile function and $p_1, p_2, \ldots, p_n$ are uniform plotting positions $(0 < p_i < 1)$. The method for calculating p_i $(i = 1, \ldots, n)$ depends on the theoretical criterion used. Common methods include: [Blom] $p_i = (i - 0.375)/(n + 0.25)$; [Weibull] $p_i = i/(n + 1)$; [Hazen] $p_i = (i - 0.5)/n$; many others are described in Aldenberg *et al.*[33] (Section 5.6.4). Looney and Gulledge[34] review hypothesis tests for assessing normality based on the Q–Q plot empirical correlation coefficient under different plotting positions. Later, in deriving formulas for regression-based estimators, we exploit the fact that all standard plotting position choices are symmetric, so that $x_i = x_{n-i+1}$ $(i = 1, \ldots, n)$.

The quantity to be estimated is $\log_{10}(\mathrm{HC}_p)$, which we denote by δ_p. For a log-normal SSD, $\delta_p = \mu - K_p\sigma$, where K_p is defined by convention as the $(100 - p)$th percentile of a standard normal distribution,[13] for example, $K_5 = 1.6449$. We will use $\hat{\delta}_p(\mathbf{Y})$ as a generic notation for rules for calculating estimates of δ_p from the data. It is foreseen that without wider modeling assumptions, the risk assessor will reuse the determined rule independently for future risk assessments.

2.1. Backwards Regression Estimator

Wheeler *et al.*[28,29] assume the following model for describing the SSD:

$$X = a + bY + \epsilon,$$

where X is a percentile in probit scale, Y is a log concentration, a is an intercept parameter, b is a slope parameter, and ϵ is the error term. The model is implemented in a U.S. Environmental Protection Agency (U.S. EPA) spreadsheet for fitting SSDs (freely downloadable from: http://cfpub.epa.gov/caddis/downloads/SSD_Generator_V1.xlt; accessed 24/06/2010). A similar model has also been considered[35] under the assumption of a log-logistic SSD. The role of X can be interpreted as a one-to-one transformation of the (potentially) affected fraction of species[36] at a given environmental log concentration Y. Assuming symmetric plotting positions, so that $\sum_{i=1}^{n} x_i = 0$, simple linear regression yields estimates of a and b to be:

$$\hat{a} = -\hat{b}\bar{y}; \text{ and}$$

$$\hat{b} = \frac{\sum_{i=1}^{n} y_{(i)}x_i}{(n-1)s^2},$$

where $\bar{y}$ and s^2 are the usual minimum variance unbiased estimators of μ and σ^2; $\hat{b}^{-1}$ provides an estimate of the standard deviation parameter σ. An estimate of δ_p is obtained by inverting the predictive regression line to yield:

$$\hat{\delta}_p(\mathbf{Y})_{[\mathrm{B}]} = \bar{y} - K_p\frac{(n-1)s^2}{\sum_{i=1}^{n} y_{(i)}x_i}. \qquad (1)$$

2.2. Forwards Regression Estimator

An alternative (forwards) model[26] arises by swapping the roles of the regressor and response variables and is described[1] by the U.S. EPA Water Science Standard Academy (see: http://www.epa.gov/waterscience/standards/academy/supp/aquatic/page10.htm; accessed 24/06/2010). It is important in the field of environmental data analysis in the presence of nondetects[31] and is implicitly consistent with another model[2] that assumes a log-triangular SSD and very strong *ad hoc* criteria beyond the

[1] The U.S. EPA describe this for informational purposes only. This is not an official statement of the U.S. EPA policy.

scope of this research. We write the model as:

$$Y = \alpha + \beta X + \epsilon', \tag{2}$$

where X and Y are defined, α and β are different intercept and slope parameters, and ϵ' is an error term. As for the backwards regression estimator, simple linear regression yields estimates of α and β:

$$\hat{\alpha} = \bar{y}; \text{ and}$$
$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_{(i)} x_i}{\sum_{i=1}^{n} x_i^2},$$

and $\hat{\beta}$ is an estimate of σ. An estimate of δ_p is then:

$$\hat{\delta}_p(\mathbf{Y})_{[\mathrm{F}]} = \bar{y} - K_p \frac{\sum_{i=1}^{n} y_{(i)} x_i}{\sum_{i=1}^{n} x_i^2}. \tag{3}$$

Other statistical procedures than simple least-squares regression for estimating the slope β have been considered and empirical arguments provided for choosing such alternative fitting methods.[26]

2.3. A Confidence/Credible Limit Estimator

Aldenberg and Jaworska[13] refined the way to estimate the uncertainty of the HC_p by extending the work of others[16,17,37] who had proposed using confidence limits. The idea was to control the probability that an estimator *underestimates* the true value for 100γ% of repeated experiments, that is,

$$P[\bar{y} - ks \leq \mu - K_p \sigma] = \gamma, \tag{4}$$

where k is a constant, independent of the data and the probability is taken with respect to sampling distribution of Y. Through solving this, the 100γ% one-sided confidence limit was determined to be:

$$\hat{\delta}_p(\mathbf{Y})_{[\mathrm{C}]} = \bar{y} - \frac{1}{\sqrt{n}} T_{n-1, K_p\sqrt{n}, \gamma} s, \tag{5}$$

where $T_{n-1, K_p\sqrt{n}, \gamma}$ is the 100γ-th percentile of a noncentral t-distribution with $n-1$ degrees of freedom and noncentrality parameter $K_p\sqrt{n}$; see Aldenberg and Jaworska[13] (pp. 14–15) for a derivation. Unlike the other rules discussed here, this one was derived to have the property that, for a sufficiently large number of independent repeated samples, the HC_p estimator underestimates the true value for exactly 100γ% of the samples.

It was also proved[13,33] that the frequentist coverage of the estimator matched the Bayesian credibility when the latter was determined using the independent Jeffreys prior distribution, $P(\mu, \sigma^2) \propto \sigma^{-2}$ for $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$. More precisely, Bayesian credible limits coincided with frequentist confidence limits for the same γ. Others have discussed the use of Bayesian methods in an SSD framework[38–41] and additional properties of this confidence/credible limit estimator.[14]

Current requirements under the REACH guidance document[9] for estimating the HC_5 are to set $\gamma = 0.50$ and also to calculate a confidence interval; the Bayesian viewpoint is not explicitly mentioned. In some more general situations it has been recommended[15,16,37] to estimate the HC_5 by setting $\gamma = 0.95$ in order to yield conservatively protective estimators and it has been shown[14] that this corresponds to a highly conservatively asymmetric loss profile.

2.4. An Example

A hypothetical chemical is assessed with eight species, yielding (randomly generated) toxicity values: 0.66, 2.17, 4.33, 5.00, 9.40, 11.00, 34.90, and 2,500 mg/L. Standard practice in assessment of normality is through the Anderson-Darling (AD) test and the Kolmogorov-Smirnov (KS) test (implemented through the less powerful Lilliefors test when the alternative hypothesis is not well specified, as is the case here). These two tests are recommended by ECHA [9] (p. 22) as supporting evidence for the use of SSD-based risk assessment methods. Two other tests based on plotting positions rather than the empirical distribution function are the Shapiro-Wilk (SW) test and the Shapiro-Francia (SF) test; the latter is inherently related to the regression-based models since its test statistic is the square of the Q–Q plot correlation coefficient. The four tests in this case yield p-values 0.10 (AD), 0.12 (KS), 0.12 (SW), and 0.06 (SF); none rejects log-normality at the 5% significance level and only SF at the 10% level. Graphical inspection using a Q–Q plot (shown in Fig. 1; left panel) based on Weibull plotting positions does not strongly suggest departures from log-normality given the small sample size.

For brevity, we denote the three rules above as [B] (backwards estimator), [F] (forwards estimator), and [C] (confidence estimator with $\gamma = 0.5$). Applying each rule and transforming back onto the original concentration scale yields HC_5 estimates: 0.0438 [B], 0.118 [F], and 0.162 [C] mg/L. The median of the SSD, namely, the HC_{50}, is estimated to be 11.37 mg/L by each rule since the rules coincide at $p = 50$. The

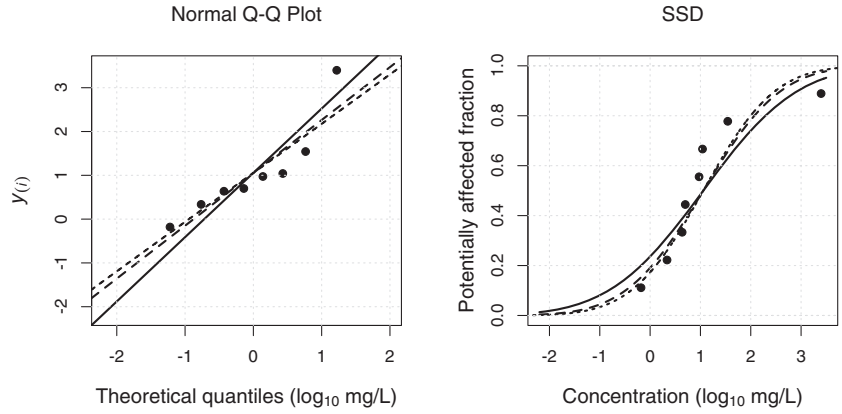

Fig. 1. Estimated SSDs based on three approaches: [B] (—), [F] (- - -), and [C] ($\cdots$). Left panel: data (points) are plotted as a Q–Q plot according to probit transformed Weibull plotting positions. [F] is directly overlaid as the forward model fitted regression line $Y = \hat{\alpha} + \hat{\beta}X$; [B] is the inverted fitted regression line $(Y = (X - \hat{a})/\hat{b})$; [C] is plotted by applying the probit transform to the estimated PAF using previously published equations.[13] Right panel: as per left panel with (x, y)-coordinates flipped and X transformed to the probability scale (using the normal cumulative distribution function).

estimates of the standard deviation σ (measured in $\log_{10}$ mg/L) are 1.47 [B], 1.21 [F], and 1.07 [C]. An overall perspective is gained from Fig. 1, which shows each fitted SSD. The left panel shows the standard Q–Q plot (data points), with the [B] (solid line), [F] (dashed line), and [C] (dotted line) probit log linearized SSDs shown. Although not apparent from the figure, the SSD for [C] is not a straight line. The right panel shows the SSDs in a more standard manner, namely, as cumulative distribution functions over log concentration. The probability measure on the vertical axis is referred to as the potentially affected fraction (of species, PAF) in the ecotoxicological literature.[36]

It is clear that the SSDs are very similar in this example; for samples with larger p-values from the goodness-of-fit test this tends to be particularly true. Nonetheless, the rules tend to diverge in the tail region where there is greatest interest for risk managers.

3. EVALUATION OF PERFORMANCE

In order to evaluate the performance of the three fixed estimator proposals [B], [F], and [C], a criterion for comparison must be specified. Recalling that the predictive problem at hand is one of estimating the HC_p, we propose measuring the performance of the estimators through either (i) direct consideration of discrepancy between the estimator and true value, or (ii) consideration of the discrepancy between the actual PAF, at the estimator concentration, and the intended PAF. The nonlinear form of SSDs is what separates these two proposals. In what follows we provide some background to each of them. There are in fact many ways to construct such criteria, for example, log-odds ratios, and they may equally well be used to derive, rather than justify, estimators *a posteriori*. This is actually a subtle, yet fundamental, issue in risk assessment.

3.1. (Standardized) Discrepancy Between log(HC_p)

A standardized measure of discrepancy[14] is defined to be $(\hat{\delta}_p - \delta_p)/\sigma$. There are two important features to this: (1) the difference is between log-transformed estimators, and (2) the difference is scaled by the log-SSD standard deviation.

One might argue that performance of a rule should be measured on the original concentration and not the log-transformed concentration scale because the rule will be (re-)used as $10^{\hat{\delta}_p}$ in the risk characterization stage of the assessment. Our argument against this position is twofold. First, [B] and [F] are derived from a statistical analysis of log concentration in order for errors (although pseudo in this case; consult Section 4.2) to be additive; the transformation back onto original scale is *post hoc*. Second, consider mapping the standardized discrepancy metric $(\hat{\delta}_p - \delta_p)/\sigma$ onto the original scale, which would yield $(\widehat{\mathrm{HC}}_p/\mathrm{HC}_p)^{1/\sigma}$, where we define $\widehat{\mathrm{HC}}_p = 10^{\hat{\delta}_p}$. The ratio $\widehat{\mathrm{HC}}_p/\mathrm{HC}_p$ clearly provides a meaningful comparison; if it is greater (or lower) than unity then the rule has under- (or over-) estimated the true value. Now, consider the following two ratios: (i) $\mathrm{HC}_{50}/\mathrm{HC}_5$ and (ii) $\mathrm{HC}_{95}/\mathrm{HC}_5$. It is straightforward to show that these ratios reduce to (i) K_5^{σ} and (ii) $K_5^{2\sigma}$, respectively. Clearly, therefore, the ratio of interest is dependent on σ. Raising the ratio to the power of $1/\sigma$ removes this dependency, consequently yielding a standardized measure (i.e.,

independent of the parameters in the SSD model) of discrepancy.[42]

3.2. Discrepancy Between log(PAF)

It has been suggested[26] that, for a given decision rule, performance could be measured with respect to the attained PAF of species affected, which is the probability that a randomly selected species from the assemblage has its toxicity value exceeded at a specified environmental concentration. We, on the other hand, propose considering the discrepancy with respect to the attained log(PAF), thus emphasizing discrepancies in a region where p is small. In other words, we measure the performance of the estimator by directly calculating the log(PAF), denoted $\log(\hat{p})$, from the true SSD (conditional upon the model parameters) evaluated at the estimated value $\hat{\delta}_p$ and subtract the intended log(PAF), $\log(p)$. Since PAF is already on a nonambiguous measurement scale (i.e., that of probability), we do not require any standardization. By considering logarithms, much greater discrepancy is assigned between, say, $p = 0.01$ and 1.0 than between $p = 0.1$ and 1.0.

3.3. Simulation

Evaluation of the three estimators is based on the following Monte Carlo simulation study. Recall that under normality, if μ and σ are known, then $\delta_p = \mu - K_p\sigma$ (and $\mathrm{HC}_p = 10^{\mu - K_p\sigma}$). For each iteration of $N = 20 \times 10^6$ simulations with a fixed specification of sample size (of toxicity data) n and protection goal p:

(1) Randomly generate n log toxicity values: $y_1, y_2, \ldots, y_n$ from a normal distribution with mean μ and standard deviation σ.
(2) Calculate: $\hat{\delta}_p(\mathbf{Y})_{[\mathrm{B}]}$ (Equation (1)), $\hat{\delta}_p(\mathbf{Y})_{[\mathrm{F}]}$ (Equation (3)), and $\hat{\delta}_p(\mathbf{Y})_{[\mathrm{C}]}$ (Equation (5)).
(3a) Calculate the standardized discrepancy between estimated and true $\log(\mathrm{HC}_p)$ for each of [B], [F], and [C].
(3b) Calculate the discrepancy between actual and intended log(PAF) for each of [B], [F], and [C].

For both discrepancy measures, the results of the study will not depend on the values of μ and σ and it suffices to fix $\mu = 0$ and $\sigma = 1$.

For each sample size n, protection goal p, and discrepancy measure we empirically summarized the Monte Carlo sample by computing (a) the mean discrepancy, (b) the median, (c) the standard deviation, (d) the root mean square, and (e) proportion of negative values. The first two summaries, (a) and (b), are the usual estimator bias and median bias, taking $\log(\mathrm{HC}_p)$ and log(PAF) as the quantities being estimated; (c) and (d), respectively, measure estimator variability and accuracy; for all four, the $\log(\mathrm{HC}_p)$ measure is standardized by σ. The final summary, (e), approximates the probability that the estimator falls below the true value and is a measure of bias for the dichotomous outcome that the estimator either under- or overestimates the true value; we call it confidence hereafter. By construction of the confidence limit rule [C] (see Equation (4)), (e) should be equal to $\gamma = 0.50$ for the $\log(\mathrm{HC}_p)$ discrepancy, a required property for an estimator to be consistent with regulatory guidance.[9] In fact, this also holds for the log(PAF) discrepancy because (e) is invariant to one-to-one transformations of the concentration scale and $\log(\mathrm{HC}_p)$ and log(PAF) are related by the SSD cumulative distribution function; this invariance also reduces the importance of the scale on which estimator rule performance is analyzed.

Different summary measures may lead to different conclusions about which estimators are "optimal." We are unaware of any omnibus criterion for measuring estimator performance and acknowledge that other criteria may be suitable. However, considering the bias and median bias as an example, it is perhaps more intuitive to use the latter. If, as we expect, risk managers treat risk assessments separately, then, despite the fact they reuse the same decision rule for consistency, they are unlikely to be interested in the mean PAF discrepancy across independent assessments as this will have no environmental interpretability. On the other hand, the median discrepancy informs the risk manager about the proportion of risk assessments that will exceed the permitted PAF.

3.4. Application

The standard policy decision regarding the fraction of species that may be permitted to be affected is $p = 5\%$ in many regulatory contexts (e.g., ECHA).[10] However, other choices such as $p = 10\%$ are of interest to scientists.[39] The need to minimize costs to industry from laboratory experimentation, combined with the ethical requirement to reduce species testing, means that the sample sizes of toxicity data are typically small. Despite the minimum sample size being $n = 10$ under the EU REACH

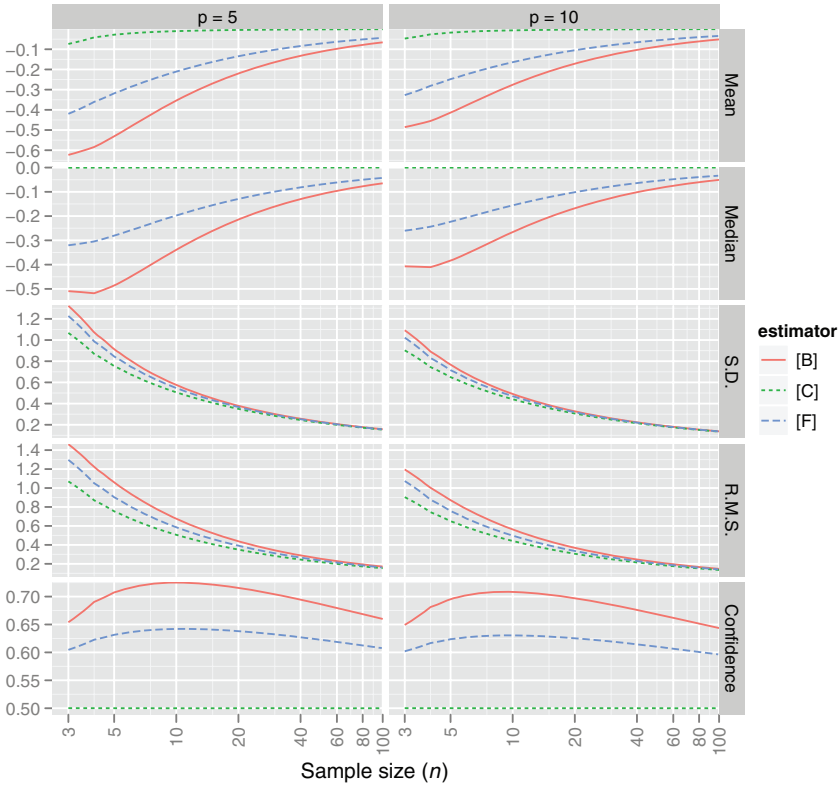

Fig. 2. Summaries of standardized discrepancy, $(\hat{\delta}_p - \delta_p)/\sigma$, for three rules $\hat{\delta}_p(\mathbf{Y})$, plotted against sample size for $p = 5$ and $p = 10$.

guidance,[10] it has been reported[33] that sample sizes much lower than 10 are not exceptional. Therefore, we evaluate the three rules, [B], [F], and [C], according to the performance criteria detailed earlier for $p = 5$ and 10 and $n = 3, \ldots, 100$. The [F] and [B] rules are determined using the Weibull-based plotting positions, which is consistent with the work of others,[2,43] and can be justified.[26] Frustratingly, not all end-users describe which plotting positions they have implemented. The results of the simulation experiment are shown in Fig. 2 (standardized discrepancy between $\log(\mathrm{HC}_p)$) and Fig. 3 (discrepancy between $\log(\mathrm{PAF})$).

4. DISCUSSION

4.1. Implications

The interpretation of Fig. 2 is straightforward: based on (standardized) $\log(\mathrm{HC}_p)$ discrepancy, the [C] estimator outperforms both [B] and [F]. For the more "plausible" regions of small n (emphasized on the figure by logarithmic horizontal axes), the differences are especially pronounced. In all three cases the mean discrepancy was negative, suggesting that the procedures are conservative on average, most noticeably in the case of [B]. Note that the mean discrepancy of the [C] estimator is nonzero because that summary measures expectation as opposed to the median discrepancy, which is zero by design for [C]; that the mean is negative is due to the left-skewed nonsymmetric sampling distribution (a rescaled noncentral t-distribution in this case) of $\hat{\delta}_p(\mathbf{Y})$. Although the larger negative mean discrepancy for [B] may appeal from a viewpoint of being protective, this is not the defined goal in estimating the HC_p.

The same conclusions are drawn when considering the discrepancy between $\log(\mathrm{PAF})$, as confirmed by Fig. 3. As stated previously, the confidence summary (panel row 5) in Figs. 2 and 3 are identical.

The level of attained confidence for the [C] estimator (see row 5 in Figs. 2 and 3) is found correctly to be 50%. In the case of [B] and [F] the most prominent feature is the lack of monotonic behavior. This inconsistency, unlike the [C] estimator, may be considered undesirable by risk managers who would intuitively seek monotonicity. Nevertheless, based on the standardized discrepancy between HC_p, the estimators demonstrate properties of conservatism since for the range of sample sizes considered, the estimators

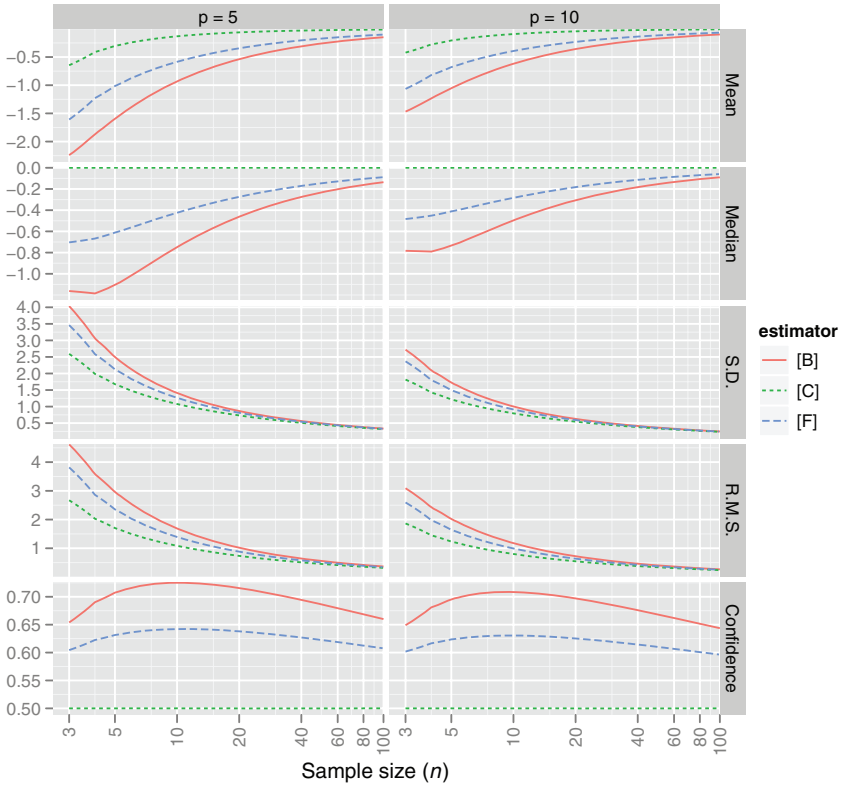

Fig. 3. Summaries of standardized discrepancy, $\log(\hat{p}) - \log(p)$, for three rules, plotted against sample size for $p = 5$ and $p = 10$.

underestimate the HC_p for between 60% and 75% of samples, with [B] consistently the most conservative.

Further analysis showed that there are paradoxical outcomes if one considers discrepancy in terms of PAF rather than log(PAF). Then the [C] estimator has the largest magnitude of mean, standard deviation, and R.M.S. discrepancy, suggesting that any decision rule determined from this criterion would lead to overachieved protection goals on average. On the other hand, however, it remains optimal from the perspective of median discrepancy due to the invariance property of the median function to one-to-one transformations.

4.2. Limitations in the Regression Methods

There are a number of further limitations to the two regression-based estimators that weaken statistical defensibility when the toxicity data sample is well defined. A few noteworthy points are discussed here.

Uncertainty quantification. If either the forward or backward model is considered tenable, then the estimators derived in Equations (1) or (3) are sound. However, if sampling uncertainty about the estimators is to be reported, the covariance matrix of ϵ (or ϵ') must be specified because the typical linear regression assumption that errors are independently and identically normally distributed (with zero mean and homogenous variance) does not apply. For a known covariance matrix, one can apply *generalized least squares*; in the case of the forward linear model (Equation (2)) this implicitly leads to the well-known SW test statistic[44] for assessing a hypothesis of normality. Calculating the appropriate covariance matrix for either linear model proposal is nontrivial. However, uncertainty quantification will be incorrect if this is not accounted for. An extensive discussion[26] included a parametric (best linear unbiased) estimator, which incorporated this covariance structure for the forward model using an approximation method. Despite this, the correct assumptions do not appear to have been incorporated elsewhere in the ecotoxicological risk assessment literature.

Dependence on plotting positions. Both estimators of $\log_{10}(HC_p)$ depend on $x_1, x_2, \ldots, x_n$ through the weighted linear sum $y_{(1)}x_1 + y_{(2)}x_2 + \cdots + y_{(n)}x_n$ and in the case of [F] through the sum of squares. Hence different choices of p_i will lead to different estimators, thus demonstrating the *artificial* nature of the proposals, i.e., we never actually "observe" the

species assemblage PAF at the environmental concentration log concentration $y_{(i)}$ to be x_i. Nevertheless, differences will likely be small unless there are outliers in the model, in which case the assumption of normality would need to be readdressed anyway. Recommendations have been made[26] regarding the choice of plotting positions.

Vertical versus horizontal least squares. Momentarily discounting the implication of the two previous issues, an obvious question is why should one choose either the forwards or backwards model over the other. The forwards and backwards estimated SSDs result from minimizing, respectively, the sum of vertical and horizontal squared differences between the log-linearized SSD and the data points in the Q–Q plot. From a standard statistical viewpoint, we consider that the forward model is more natural since the toxicity data (the response variable) are, after all, what we consider to be randomly distributed, whereas the transformed plotting positions (the regressors) are supposedly known; we are not alone in this view.[26] Furthermore, this is consistent with the application of such procedures in environmental data analysis in the presence of censored data.[31] A related issue with the backwards model for estimation of δ_p is that under this model we actually want to predict the regressor from the response. This so-called linear calibration problem is a well-studied topic in the statistical literature[45] with no definitive solution.

5. CONCLUSIONS

There is a clear confusion in the SSD literature regarding the term "parametric SSD," despite literature[26] differentiating between parametric and so-called graphical approaches. In one case, authors[29] cite another article[16] that is effectively based on the confidence limit approach and yet implicitly use a so-called linearized log-normal method, what we call the backwards regression estimator. Such situations are confusing to SSD practitioners who may lack the theoretical knowledge to differentiate between the two approaches, let alone between the consequences of the decision rules that result.

Regression-based estimation methods are popular in ecotoxicological and environmental risk (hazard) assessment; early proposals in a regulatory context were introduced by the U.S. EPA.[2,26] We do not altogether reject their application since they are transparent compared to the fixed assessment factor approach. Moreover, their use in certain cases is beyond rigorous statistical reasoning; in one example[2] a forward regression model is used, but only through the four toxicological endpoint measurements nearest to 5th percentile; such *ad hoc* behavior, despite attempted empirical reasoning,[26] lacks statistical justification. Q–Q plots, from which such estimators effectively derive, are indeed useful, if not altogether necessary, for assessing the validity of (log-) normality. A key advantage of regression methods discussed here is for estimation of location and scale in the presence of censored data; however, such data are regularly not accepted in regulatory risk assessment. Moreover, for the special case of log-normality, the statistical problem of percentile estimation for censored data is well developed,[31,46] reducing the need for regression-based estimators due to the ready availability of statistical computing software. The situation is more difficult for other distribution families; it has been noted[26] that for alternative distributions, "parametric" methods may be increasingly difficult to analyze in comparison to "graphical" methods. Despite the more widespread use of the backwards regression model, its statistical foundation lacks credibility compared to the forwards model.

If the primary current motivation for regression-based estimators is, as we suspect, the ease of implementation by practitioners (since regression software is readily available and simple to use), then it is perhaps misunderstood how straightforwardly the standard confidence limit approaches can be used. Statistical tables have been published,[13,33] which can be quickly used to determine HC_p estimators and corresponding confidence intervals; this procedure was made available[47] via a freely available software package—$E_T X$ 2.0 (http://www.rivm.nl/rvs/risbeoor/Modellen/ETX.jsp). If simple crude estimates are what in fact practitioners seek, then an even simpler approach would be to adopt, as some have,[32] a method-of-moments approach where an assessor uses the standard unbiased estimates of the location and scale parameters (in this instance $\bar{y}$ and s^2) as plug-in values for the unknown true values. Under log-normality, this would yield $\hat{\delta}_p(\mathbf{Y}) = \bar{y} - K_p s$. No measure of uncertainty is attached to this estimator, but no easily determinable correct measure of uncertainty is attached to the regression-based estimators as discussed in Section 4.2.

Debates about the application of parametric SSDs are ongoing,[15,19,39] but they remain peripheral to our intention, which is to provide purely statistically based recommendations to support

practitioners. We recommend that (1) where a pointwise estimate of the HC_p is required and the available toxicity sample is noncensored, the direct confidence limit estimation approaches[13,16,17] should be used rather than the regression-based estimators; and (2) pointwise estimators should complement, not replace, a probabilistic distribution.

Our first recommendation comes directly from the analysis in this research. However, this should not preclude "back of envelope" calculations, for example, moment-based estimators, for purely exploratory (nonregulatory) risk assessment. The second recommendation is a by-product of our analysis: since there is no "best" choice of estimator without a well-defined criterion, either a choice should be justified *a priori* or a full probabilistic description of uncertainty concerning the HC_p and/or SSD should be considered in each risk assessment. The Bayesian paradigm is natural for this and has gained considerable recent attention.[13,14,38–40,48] In fact, it has been shown[13] that the (median) [C] estimator discussed in this research is the median of the posterior distribution of the HC_p using a standard noninformative prior distribution. Furthermore, handling censored data is simple in a Bayesian framework, thus allowing for consistency in risk assessment approaches.

In principle, (1) should be easy to impose within modern regulatory frameworks; it is already part of REACH guidance. However, (2) is challenging, especially to risk assessment frameworks still using "assessment factor" based estimators, which lack a statistical basis and provide no measure of uncertainty pertinent to the safety assessment. Recent projects such as EUFRAM[7] have called for further application of probabilistic techniques in ecotoxicological risk assessment. To justify a single estimator requires a measure of loss for all possible discrepancies. The decision-theoretic concept of loss functions has been used[14,41,49] to derive HC_p estimators and it has been concluded[14] that for defensible *conservative* estimators, a loss-function approach is ideal; however, it would be likely to perplex most risk managers due to the sophisticated quantitative reasoning required.

Our discussion has been limited to the log-normal SSD because of its prevalence in the REACH guidance,[9] but we expect the results to hold for other distribution families; for example, corresponding confidence-based estimators have been derived[17] for the log-logistic SSD. The (normal) Q–Q plot, which is inherently associated with the two regression-based estimators, should still be used as a diagnostic device for assessing distributional assumptions, although the strengths of this and the statistical power of goodness-of-fit tests are severely limited at very small sample sizes.

ACKNOWLEDGMENTS

We thank all those who made suggestions about this article, including: Stuart Marshall, Peter Chapman, and Oliver Price (Unilever), Mathijs Smit (Statoil ASA), Andy Hart (the Food and Environment Research Agency), Malyka Galay-Burgos (ECETOC), Robert Luttik and Tom Aldenberg (RIVM), and the anonymous reviewers. We also thank Unilever and Statoil ASA who funded this research as part of a wider project into the study of SSDs and PNECs.

REFERENCES

1. Posthuma L, Suter GW II, Traas TP (eds). Species Sensitivity Distributions in Ecotoxicology. Boca Raton: Lewis Publishers, 2002.
2. Stephan CE, Mount DI, Hansen DJ, Gentile JH, Chapman GA, Brungs WA. Guidelines for Deriving Numerical National Water Quality Criteria for the Protection of Aquatic Organisms and Their Uses. Duluth, MN: U.S. Environmental Protection Agency Office of Research and Development Report, PB 85-227049, 1985.
3. ECOFRAM (Ecological Committee on FIFRA Risk Assessment Methods). ECOFRAM Aquatic Draft Report, 1999. Available at http://www.epa.gov/oppefed1/ecorisk/aquareport.pdf, Accessed on September 1, 2011.
4. ECOFRAM (Ecological Committee on FIFRA Risk Assessment Methods). ECOFRAM Terrestrial Draft Report, 1999. Available at http://www.epa.gov/oppefed1/ecorisk/terrreport.pdf, Accessed on September 1, 2011.
5. ANZECC and ARMCANZ (Australian and New Zealand Environment and Conservation Council and Agriculture and Resource Management Council of Australia and New Zealand). Australian and New Zealand Guidelines for Fresh and Marine Water Quality. Canberra: ANZECC and ARMCANZ National Water Quality Management Strategy Paper, No. 4, 2000.
6. EFSA (European Food Safety Authority). Opinion of the Scientific Panel on Plant Health, Plant Protection Products and Their Residues (PPR) on a request from EFSA related to the assessment of the acute and chronic risk to aquatic organisms with regard to the possibility of lowering the uncertainty factor if additional species were tested. EFSA Journal, 2006; 301:1–45.
7. EUFRAM. The EUFRAM Framework: Introducing Probabilistic Methods into the Ecological Risk Assessment of Pesticides, Volume of the EUFRAM Project Report, Volume 1, Version 5, 2006. Available at: http://www.eufram.com/documents/EUFRAM.
8. CCME (Canadian Council of Ministers of the Environment). A Protocol for the Derivation of Water Quality Guidelines for the Protection of Aquatic Life. Winnipeg, MB: CCME, 2007.

9. ECHA (European Chemicals Agency). Guidance for the Implementation of REACH: Guidance on Information Requirements and Chemical Safety Assessment, Chapter R.10: Characterisation of Dose [Concentration]-Response for Environment, May 2008. Available at: http://guidance.echa.europa.eu/docs/guidance_document/information_requirements_r10_en.pdf, Accessed on September 1, 2011.
10. ECHA (European Chemicals Agency). Guidance for the Implementation of REACH: Guidance on Information Requirements and Chemical Safety Assessment, Chapter R.19: Uncertainty Analysis, May 2008. Available at http://guidance.echa.europa.eu/docs/guidance_document/information_requirements_r19_en.pdf, Accessed on September 1, 2011.
11. Office for the Coordination of Humanitarian Affairs, United Nations. The Flash Environmental Assessment Tool (FEAT): To identify acute environmental risks immediately following disasters, Version 1.1. Geneva: United Nations, OCHA/ESB/2009/16, August 2009.
12. Van Straalen NM. Threshold models for species sensitivity distributions applied to aquatic risk assessment for zinc. Environmental Toxicology and Pharmacology, 2002; 11:167–172.
13. Aldenberg T, Jaworska JS. Uncertainty of the hazardous concentration and fraction affected for normal species sensitivity distributions. Ecotoxicology and Environmental Safety, 2000; 46:1–18.
14. Hickey GL, Craig PS, Hart A. On the application of loss functions in determining assessment factors for ecological risk. Ecotoxicology and Environmental Safety, 2009; 72:293–300.
15. Newman MC, Ownby DR, Mezin LCA, Powell DC, Christensen TRL, Lerberg SB, Anderson B-A. Applying species sensitivity distributions in ecological risk assessment: Assumptions of distribution type and sufficient numbers of species. Environmental Toxicology and Chemistry, 2000; 19:508–515.
16. Wagner C, Løkke H. Estimation of ecotoxicology protection levels from NOEC toxicity data. Water Research, 1991; 25:1237–1242.
17. Aldenberg T, Slob W. Confidence limits for hazardous concentrations based on logistically distributed NOEC toxicity data. Ecotoxicology and Environmental Safety, 1993; 25: 48–63.
18. Shao Q. Estimation for hazardous concentrations based on NOEC toxicity data: An alternative approach. Environmetrics, 2000; 11:583–595.
19. Forbes VE, Calow P. Species sensitivity distributions revisited: A critical appraisal. Human and Ecological Risk Assessment, 2002; 8:1625–1640.
20. Kefford BJ, Palmer CG, Jooste S, Warne M, Nugegoda D. "What is it meant by 95% of species"? An argument for the Inclusion of rapid tolerance testing. Human and Ecological Risk Assessment, 2005; 11:1025–1046.
21. Staples CA, Woodburn KB, Klecka GM, Mihaich EM, Hall AT, Ortego L, Caspers N, Hentges SG. Comparison of four species sensitivity distribution methods to calculate predicted no effect concentrations for Bisphenol A. Human and Ecological Risk Assessment, 2008; 14:455–478.
22. Baird DJ, Van den Brink PJ. Using biological traits to predict species sensitivity to toxic substances. Ecotoxicology and Environmental Safety, 2007; 67:296–301.
23. Kooijman SALM. A safety factor for LC50 values allowing for differences in sensitivity among species. Water Research, 1987; 21:269–276.
24. Van Straalen NM, Van Leeuwen CJ. European history of species sensitivity distributions. Pp. 19–34 in Posthuma L, Suter GW II, Traas TP (eds). Species Sensitivity Distributions in Ecotoxicology. Boca Raton: Lewis Publishers, 2002.
25. Rice JA. Mathematical Statistics and Data Analysis, 2nd ed. Belmont, CA: Duxbury Press, 1995.
26. Erickson RJ, Stephan CE. Calculation of the Final Acute Value for Water Quality Criteria for Aquatic Organisms. Duluth, MN: U.S. Environmental Protection Agency, Office of Research and Development Report, PB 88-214994, 1988.
27. Newman MC, Ownby DR, Mezin LCA, Powell DC, Christensen TRL, Lerberg SB, Anderson B-A, Padma TM. Species sensitivity distributions in ecological risk assessment: Distribution assumptions, alternate bootstrap techniques, and estimation of adequate number of species. Pp. 119–132 in Posthuma L, Suter GW II, Traas TP (eds). Species Sensitivity Distributions in Ecotoxicology. Boca Raton: Lewis Publishers; 2002.
28. Wheeler JR, Grist EPM, Leung KMY, Morritt D, Crane M. Species sensitivity distributions: Data and model choice. Marine Pollution Bulletin, 2002; 45:192–202.
29. Wheeler JR, Leung KMY, Morritt D, Sorokin N, Rogers H, Toy R, Holt M, Whitehouse P, Crane M. Freshwater to saltwater toxicity extrapolation using species sensitivity distributions. Environmental Toxicology and Chemistry, 2002; 21:2459–2467.
30. Kwok KWH, Leung KMY, Lui GSG, Chu VKH, Lam PKS, Morritt D, Maltby L, Brock TCM, Van den Brink PJ, Warne M, Crane M. Comparison of tropical and temperate freshwater animal species acute sensitivities to chemicals: Implications for deriving safe extrapolation factors. Integrated Environmental Assessment and Management, 2007; 3:49–67.
31. Helsel DR. Nondetects and Data Analysis: Statistics for Censored Environmental Data. NJ: Wiley, 2005.
32. De Zwart D. Observed regularities in SSDs for aquatic species. Pp. 133–154 in Posthuma L, Suter GW II, Traas TP (eds). Species Sensitivity Distributions in Ecotoxicology. Boca Raton: Lewis Publishers, 2002.
33. Aldenberg T, Jaworska JS, Traas TP. Normal species sensitivity distributions and probabilistic ecological risk assessment. Pp. 49–102 in Posthuma L, Suter GW II, Traas TP (eds). Species Sensitivity Distributions in Ecotoxicology. Boca Raton: Lewis Publishers, 2002.
34. Looney SW, Gulledge TR. Use of the correlation coefficient with normal probability plots. American Statistician, 1985; 39:75–79.
35. Parkhurst BR, Warren-Hicks W, Cardwell RD, Volosin J, Etchison T, Butcher JB, Covington SM. Aquatic Ecological Risk Assessment: A Multi-Tiered Approach. Alexandria, VA: Water Environment Research Agency, 1996.
36. Traas TP, Van de Meent D, Posthuma L, Hamers T, Kater BJ, De Zwart D, Aldenberg T. The potentially affected fraction as a measure of ecological risk. Pp. 315–344 in Posthuma L, Suter GW II, Traas TP (eds). Species sensitivity distributions in ecotoxicology. Boca Raton: Lewis Publishers, 2002.
37. Van Straalen NM, Denneman CAJ. Ecotoxicological evaluation of soil quality criteria. Ecotoxicology and Environmental Safety, 1989; 18:241–251.
38. Grist EPM, O'Hagan A, Crane M, Sorokin N, Sims I, Whitehouse P. Bayesian and time-independent species sensitivity distributions for risk assessment of chemicals. Environmental Science and Technology, 2006; 40:395–401.
39. Hickey GL, Kefford BJ, Dunlop JE, Craig PS. Making species salinity sensitivity distributions reflective of naturally occurring communities: Using rapid testing and Bayesian statistics. Environmental Toxicology and Chemistry, 2008; 27:2403–2411.
40. Fox D. A Bayesian approach for determining the no effect concentration and hazardous concentration in ecotoxicology. Ecotoxicology and Environmental Safety, 2010; 73:123–131.
41. Hickey GL. Ecotoxicological risk assessment: Developments in PNEC estimation. PhD thesis, Durham University, 2010.
42. Zieliński R. Estimating quantiles with Linex loss function: Applications to VaR estimation. Applicationes Mathematicae, 2005; 32:367–373.

43. Solomon KR, Takacs P. Probabilistic risk assessment using species sensitivity distributions. Pp. 285–313 in Posthuma L, Suter GW II, Traas TP (eds). Species Sensitivity Distributions in Ecotoxicology. Boca Raton: Lewis Publishers, 2002.
44. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika, 1965; 52:591–611.
45. Osborne C. Statistical calibration: A review. International Statistical Review, 1991; 59:309–336.
46. Helsel DR. Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. Chemosphere, 2006; 65:2434–2439.
47. Van Vlaardingen PLA, Traas TP, Wintersen AM, Aldenberg T. ETX 2.0: A program to calculate hazardous concentrations and fraction affected, based on normally distributed toxicity data. Bilthoven, The Netherlands: RIVM Report, 601501028/2004, 2004.
48. Hayashi TI, Kashiwagi N. A Bayesian method for deriving species-sensitivity distributions: Selecting the best-fit tolerance distributions of taxonomic groups. Human and Ecological Risk Assessment, 2010; 16:251–263.
49. Chen L. A conservative, non-parametric estimator for the 5th percentile of the species sensitivity distribution. Journal of Statistical Planning and Inference, 2003; 123:243–258.