PERGAMON

Marine Pollution Bulletin 45 (2002) 192–202

MARINE POLLUTION BULLETIN

www.elsevier.com/locate/marpolbul

Species sensitivity distributions: data and model choice

J.R. Wheeler[a,*], E.P.M. Grist[a], K.M.Y. Leung[a,b], D. Morritt[a], M. Crane[a]

[a] School of Biological Sciences, Royal Holloway, University of London, Egham, Surrey TW20 OEX, UK
[b] The Swire Institute of Marine Science and Department of Ecology and Biodiversity, The University of Hong Kong, Cape d'Aguilar, Shek O, Hong Kong

Abstract

Species sensitivity distributions (SSDs) are increasingly incorporated into ecological risk assessment procedures. Although these new techniques offer a more transparent approach to risk assessment they demand more and superior quality data. Issues of data quantity and quality are especially important for marine datasets that tend to be smaller (and have fewer standard test methods) when compared with freshwater data. An additional source of uncertainty when using SSDs is appropriate selection from the range of methods used in their construction. We show through examples the influence of data quantity, data quality, and choice of model. We then show how regulatory decisions may be affected by these factors.

Keywords: Ecological risk assessment; Species sensitivity distribution; Data quantity; Data quality; Model fit

1. Introduction

Species sensitivity distributions (SSDs) are increasingly used in ecological risk assessment procedures (e.g., Solomon et al., 1996; Steen et al., 1999) and formulation of water quality guidelines (ANZECC and ARMCANZ, 2000). This is because, when used correctly, they can introduce greater statistical confidence into risk assessment processes when compared to traditional quotient and assessment factor approaches. In Europe, risk assessment methods for new and existing chemicals are described in the technical guidance document (TGD) developed by the European Commission, the European Union member states and the European Chemical Industries (Crane et al., 2001). The TGD is currently under revision, and the inclusion of statistical extrapolation methods using SSDs is likely to be adopted.

The aim of a SSD analysis is to determine a chemical concentration protective of most species in the environment. Usually a point estimate known as the HC5 (hazardous concentration for 5% of species), or the 95% protection level (van Straalen and van Rijn, 1998) is calculated. This is a concentration that will exceed no more than 5% of species effects levels, usually based on chronic no observed effect concentrations (NOECs). It has been proposed that the lower confidence interval of the HC5, possibly with an additional safety factor of up to 10, be used to derive predicted no effect concentrations (PNECs) for risk assessment (Feibicke and Ahlers, 2001). SSDs are constructed using a cumulative plot of logarithmically transformed NOECs against rank assigned percentiles for each value to which a statistical distribution is fitted. In Europe and the United States this is typically a log-normal (Wagner and Lokke, 1991) or log-logistic (Aldenberg and Slob, 1993) model, whilst in Australia and New Zealand the Burr Type III method is used (Shao, 2000). From each of these models the HC5 endpoint is extrapolated.

To date most published ecological risk assessments using SSDs have centred on freshwater environments for which there is an abundance of good quality data, predominantly for pesticides (Solomon et al., 1996; Giesy et al., 1999; Campbell et al., 2000). There are generally fewer data available for saltwater species than for freshwater species, especially for organic compounds (e.g., Solbe et al., 1993), which presents fundamental problems when attempting to apply the SSD approach to ecological risk assessments for substances in marine environments (Leung et al., 2001). This raises questions about how much data is enough, what effect data quality

* Corresponding author. Tel.: +44-1784-414196; fax: +44-1784-470756.
E-mail address: j.wheeler@rhul.ac.uk (J.R. Wheeler).

0025-326X/02/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S0025-326X(01)00327-7

has on the overall result, and whether different methods are more suitable for analysing smaller datasets.

2. Data quantity and quality

The usefulness of SSD analyses is likely to depend, at least in part, upon the quality of input data. Inclusion of poor data will compound the problems of interpreting 'natural' variance, and will probably generate bad predictions. But how bad? SSDs are effectively statistical extrapolation techniques that require a minimum, although unknown, amount of data in order to produce reliable estimates upon which regulatory decisions may be based. We enlarged the ECETOC (European Centre for Ecotoxicology and Toxicology of Chemicals) database (Solbe et al., 1998) and adapted it to include a data quality criteria classification. Data were quality assessed according to criteria in Box 1. The SSDs were then constructed using acute data to examine the effects of dataset size.

Box 1. Data quality criteria used to assess data included in the ECETOC database.

QA1 Study of highest reliability
Studies assigned a reliability score of 1 meet all of the following criteria:

- Published or well-documented procedures cited
  ⇒ *If a 'standard' method is given, it should be assumed to have been followed unless indicated to the contrary. When a 'non-standard' method is cited, the reference should be obtained for assessment*
- Control performance (including solvent control) reported and satisfactory
- Measured concentrations (at least t_0 and t_{end}). Endpoints calculated in terms of geometric mean of the measured concentrations (or nominal if measured concentrations are within 20% of nominal)

QA2 (a) Study of moderate reliability
Studies assigned a score of 2 (a) are characterised by features such as:

- Only the concentrations of stock solutions were measured (but on the basis of physical and chemical properties of test substance no loss would be expected)
- Description of the methodology incomplete
- Control mortality not reported
- No solvent control when solvent used

QA2 (b) Study of moderate reliability
Studies assigned a score of 2 (b) are characterised by features such as:

- Toxicant concentrations were not measured (but on basis of physical and chemical properties of test substance no loss would be expected)
- Description of the methodology incomplete
- Control mortality not reported
- No solvent control when solvent used

QA3 Study of limited reliability
Studies assigned a score of 3 have limited reliability but are still included in the database. They are characterised by features such as:

- Method section shows weaknesses in experimental procedures
- Unacceptable control (and/or for solvent) mortality
- As a rule of thumb the following should be followed for assessing acceptability of control mortality:
  ⇒ *egg or early life stage tests—accept control mortality* $\leqslant 40\%$
  ⇒ *>10 organisms/concentration: accept* $\leqslant 10\%$ *control mortality* (*according to OECD criteria*)
- $\leqslant$ *10 organisms* (*minimum 7/concentration*): *accept 1 death* (*according to OECD criteria*). Alternatively, test classed unacceptable for any other reason on the basis of scientific judgement

QA4 Study unreliable
Studies assigned a score of 4 do not meet criteria for reliability and were not included in the database. They may be characterised by features such as:

- Toxicant concentration not measured and on the basis of physical and chemical properties of test substances losses would be expected
- Endpoint extrapolated beyond range of concentrations tested
- Endpoint greater than the limit of solubility

The effect of data quantity was assessed using a resampling approach in which resamples were drawn randomly (from a uniform distribution) without replacement, in sizes ranging from four to the original sample size of n. For each resample, 100 replicates were generated and curves were fitted using both the log-logistic and log-normal distributions. Regression parameters, their mean values and associated standard deviations were recorded at each resample size. Comparable results were found for both model descriptions. Here, for conciseness, results for the log-logistic distributions are given. Plots of log-logistic regression parameters, α values (location), β values (scatter) and coefficients of determination (r^2) were used to establish the minimum number of data at which stabilisation of the respective parameter value occurred. Stabilisation is

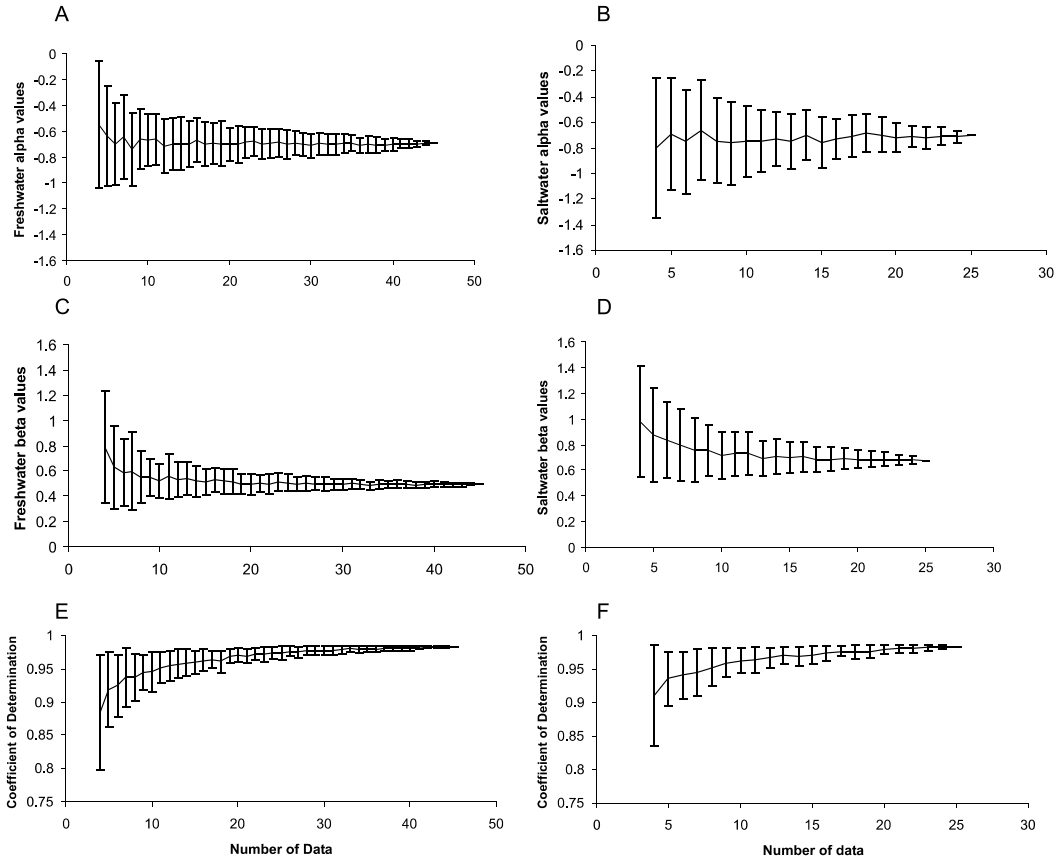

Fig. 1. Influence of data inclusion on calculation of regression parameters for the copper log-logistic distributions. The solid line connects parameter mean values computed over 100 replicates for each sample size, vertical bars show the associated standard deviations. Both freshwater and saltwater α (A and B), β (C and D) and coefficient of determination (E and F) values stabilise after 10 data points.

achieved where parameter standard deviations converge with increasing sample size. Fig. 1 shows the outcomes for acute data for copper. These plots show that stabilisation occurred at a sample size of 10–15 data points. Stabilisation profiles of pentacholorophenol were also achieved using at least 10 data points. Below 10 data points the parameter values varied widely, and would not yield a reliable estimate of a particular endpoint (e.g. the HC5). At least for these example datasets, 10–15 randomly selected data points were sufficient to obtain an effective estimation of the sub-population of data available to us (which we *assume* is representative of the total population). A minimum of data for 10 species suggested here is also in good agreement with other authors, for aquatic risk assessment (Solomon et al., 1996) and soil quality objectives (Vega et al., 1999). Although smaller datasets ($n < 10$) are often used (Aldenberg and Slob, 1993; Hakanson, 1995) this would be at the point of greatest variability in model output and may well produce unreliable estimates for specified effect levels.

The influence of data quality on resulting HC5 values was investigated by removing certain qualities of data from analyses. Removal of data in blocks, according to quality, has the confounding effect of removing particular species. This is because certain types of toxicity experiments, such as the *Daphnia magna* 48-h LC50 test, following standard protocols nearly always fell into quality category 1 while, in contrast, tests with non-standard test method species tend to fall into lower quality categories. In order to account for this, we generated expected HC5 values based on the number of species present. This was achieved by adapting a method described by Vega et al. (1999) whereby SSD outputs (e.g., HC5) were recalculated with a progressive increase in the number of data included in each analysis. Randomised removal of data enabled a range of values for the new number of species (after removal due to quality) to be calculated. Employing this method, we examined the effect of data quality on the SSD outputs for acute freshwater copper data, and showed that there was a two-fold change in the HC5 value between using data of all qualities to only using QA 1 (the best quality), yielding HC5 values of 0.0068 and 0.011 mg/l respectively (Table 1). Although the difference is of marginal significance, it is not likely to be a function of the number of species, as the expected HC5 is 0.007 mg/l. Consequently we may conclude that more stringent criteria for data inclusion would lead to a

Table 1
Effect of removing data of different qualities (derived from the Quality Assured criteria)

	QA 1, 2(a), 2(b), 3	QA 1, 2(a), 2(b)	QA 1, 2(a)	QA 1
Copper				
n	45	40	29	29
α	−0.5572	−0.5934	−0.7386	−0.6965
SD	0.9954	1.0317	0.8210	0.7755
β	0.54747	0.567435	0.45155	0.426525
r^2	0.9493	0.9414	0.9446	0.9500
HC5 (mg/l)	0.0068	0.0054	0.0086	0.011
Expected HC5	–	0.0062 ± 0.0009	0.0070 ± 0.0014	0.0070 ± 0.0014
Pentachlorophenol				
n	39	36	–	31
α	−0.8184	−0.9852	–	−1.1202
SD	1.1158	1.1336	–	1.1517
β	0.61369	0.62348	–	0.633435
r^2	0.8929	0.8496	–	0.8490
HC5 (mg/l)	0.0024	0.0015	–	0.0010
Expected HC5	–	0.0021 ± 0.0003	–	0.0025 ± 0.0003

n is the number of species, α (location) and β (scatter) parameters of the log-logistic regression, SD standard deviation of log toxicity values, HC5 (mg/l). The expected HC5 represents a value from the same number of species with all data qualities included, repeated for five different random combinations of the data, plus or minus the standard deviation of these values.

higher concentration protection level in the case of copper. However, the opposite was observed for pentachlorophenol where HC5 results varied between 0.0024 (all data) and 0.001 mg/l (QA1 data only) (Table 1). There is no a priori reason to expect that there should be a relationship between sensitivity and quality. It is, however, worth noting that where the test endpoint is a NOEC, there is evidence to indicate that poorly designed studies can yield higher NOEC estimates because frequently there is low statistical power in the significance test (Crane and Newman, 2000).

3. Distribution construction

Ecotoxicology databases often contain multiple entries for the same species and chemicals, because several toxicity tests have been performed. Risk assessors must then choose how best to summarise such multiple data when constructing SSDs. Cumulative plots of species responses are used, and because multiple data for a species exist the 'total' species response is usually estimated by calculating a mean from reliable test endpoints. However, there are four possible options for incorporating multiple species values in SSD analyses:

1. include only the single most sensitive value for each species;
2. include all reliable available data;
3. geometric mean species summaries; or
4. arithmetic mean species summaries.

Here we investigate the effect of each of these multiple species summaries. Acute saltwater lethality data for six substances were extracted from the US-EPA AQUIRE aquatic toxicology database (http://www.epa.gov/ecotox/). In all cases the log-logistic distribution proved to be a good model fit and for consistency was used throughout. Four SSDs were constructed for each substance using the most sensitive endpoint, all data, and geometric or arithmetic means. We summarised and compared the results of these analyses by using HC5 point estimates with associated lower 95% confidence intervals (one tailed) (Table 2). As we might expect, there was a general trend in HC5s and their left sided confidence intervals: the most sensitive endpoint

Table 2
HC5 values (μg/l) calculated from saltwater SSDs with four multiple data summaries for species

Substance	Most sensitive	Geometric mean	Arithmetic mean	All values
Cadmium	6.592 (0.978)	8.220 (1.070)	8.337 (1.102)	16.56 (8.476)
Chlordane	0.532 (0.041)	0.562 (0.036)	0.570 (0.039)	0.638 (0.134)
Dieldrin	0.200 (0.039)	0.817 (0.060)	0.931 (0.222)	0.783 (0.481)
Endosulfan	0.007 (0.0005)	0.014 (0.001)	0.017 (0.001)	0.040 (0.017)
Nickel	375.0 (15.01)	529.7 (25.20)	563.6 (35.56)	615.2 (63.66)
Toluene	2818 (95.31)	4271 (180.2)	4519 (189.4)	2864 (718.0)

Lower left-sided 95% confidence limit (one tailed) calculated according to Aldenberg and Slob, 1993 in parentheses.

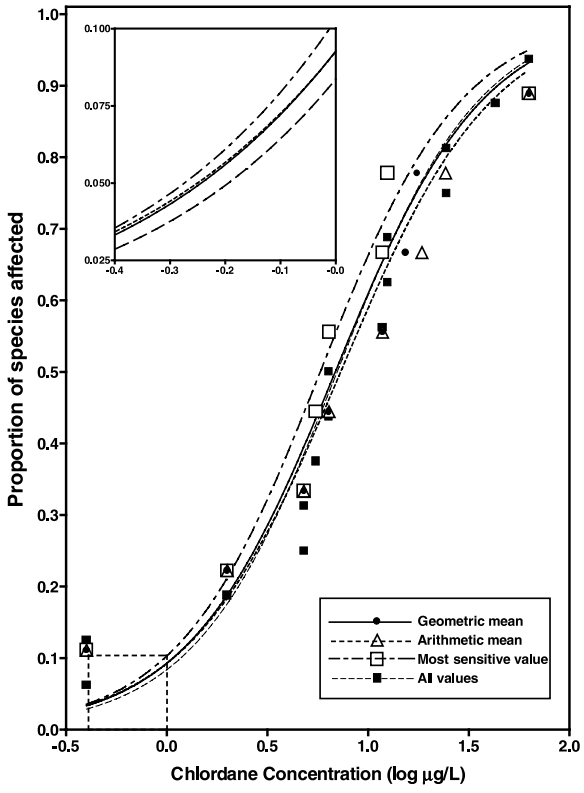

Fig. 2. Species sensitivity distribution for chlordane illustrating different positions of curves depending on which data summary is used. Inserted figure amplifies the lower percentile region. Where symbols overlap, the graphical sequence is: closed circle, open triangle, open square, closed square.

approach yielded the lowest value. Thereafter it was followed by the geometric mean, the arithmetic mean and then all values (Table 2). More importantly, values could differ by a factor of two or more (nickel, toluene, cadmium and endosulfan). However, for both dieldrin and toluene inclusion of all values provided a more conservative HC5 than mean summaries (Table 2). This is a direct result of the data points influencing the position of the curve fitted; Fig. 2 illustrates this for chlordane. A preponderance of low effect data is liable to shift the fitted curve to the left by virtue of the greater number of points plotted. For the most sensitive data or mean species summaries this is less likely to occur.

Analyses using only the most sensitive values in the dataset are an attractive regulatory option as, by definition, they are very conservative. However such an approach does not use all of the available data and is consequently more likely to be biased by outliers. The most sensitive species value may also discourage further data generation, as the larger the dataset the greater the probability that it will include extreme (low) values (Smith and Cairns, 1993). Including all the available data, capturing inter- and intra-species variation in response to a substance in one analysis, could resolve this. Operationally this would entail assigning a percentile to individual data points. However, this complicates the interpretation of the analysis and would require a new definition of the HC5, as such measures are no longer at the species level. This is contrary to the purpose of the SSD, namely to derive a protection level for a specified percentage of species (Toll et al., 2001). Therefore we are left with more traditional data summaries that, at least, are influenced by every data point but still keep the currency of species response. The geometric mean, despite recent criticism (Parkhurst, 1998), is often used as a summary statistic for aquatic toxicity data (Stephan et al., 1985). It is the back-transformed mean of a logarithmically transformed variable, and gives a conservative estimate due to the curvature of the logarithmic function. This conservative property is desirable for risk management decisions, so its use is recommended by the Ecological Committee on FIFRA Risk Assessment Methods, or ECOFRAM (www.epa.gov/oppefed1/ecorisk/aquareport.pdf). The geometric mean will always be either smaller than or equal to the arithmetic mean (Streiner, 2000). The arithmetic mean assumes that the differences between values are linear, and not a non-linear function as in the geometric mean. In addition it has been demonstrated, at least for log-normal data, that the mean squared errors (a measure of bias and variance of an estimator) are larger for arithmetic compared to geometric means (Smothers et al., 1999).

4. Model choice

Several approaches to forming SSD analyses have evolved over recent years. Differences between these approaches lie in the choice of underlying distribution such as the log-normal (Wagner and Lokke, 1991), log-logistic (Aldenberg and Slob, 1993) or Burr Type III (Shao, 2000). However, some authors argue that there is no reason to assume, or the ability to verify in small datasets, an underlying distribution for species sensitivities (Smith and Cairns, 1993; Forbes and Forbes, 1993). An alternative resampling (bootstrap) technique has been proposed (Jagoe and Newman, 1996), which does not rely on any assumed distribution. More recently a combined bootstrap regression approach has been developed, by smoothing the bootstrap with an assumed distribution, such as the log-logistic (Grist et al., in press).

4.1. Parametric methods

The linearised log-normal method has become the approach tentatively adopted by regulators following its use in several high profile probabilistic risk assessments for pesticides (Solomon et al., 1996; Giesy et al., 1999; Hall et al., 2000). Its major advantage is mathematical simplicity, because simple least squares regression can be applied to probit and log transformed data, and

confidence intervals can be calculated from assumptions of the normal distribution. If, in addition, chemical monitoring data or exposure estimates are available, an exceedence or joint probability curve (JPC) may be constructed (e.g., Giesy et al., 1999). The simple linear curves allow the JPC combination to communicate risk clearly and allow fairly simple assessment of the implications of different environmental management decisions. However, the log-normal distribution may also be limited by its simplicity, as it has been shown that 15 out of 30 datasets tested failed conformity tests for the log-normal distribution (Newman et al., 2000). It has also been suggested that datasets may contain several sub-distributions, possibly related to taxonomic differences in sensitivity (Newman et al., 2000) that may not be adequately described by a single straight line.

The log-logistic distribution, as noted earlier, generally provides a good fit to SSD data. The log-logistic function has extended tails and therefore has built-in conservatism (Aldenberg and Slob, 1993). Mathematically it is more complex than the log-normal model, especially when calculating confidence intervals. Extrapolation factors used to calculate confidence intervals have been derived through Monte Carlo simulation, and tabulated (Aldenberg and Slob, 1993). The factors, however, are restricted to a one tailed 95% interval (equivalent to two tailed 90% interval), whereas we are often interested in confidence at the two tailed 95% level. Also, for sample sizes between tabulated values linear interpolation must be used to determine an approximation of the factor.

4.2. Non-parametric bootstrap method

The application of the bootstrap to SSD analyses was proposed by Jagoe and Newman (1996) as an alternative to fitting a distribution to data. Bootstrapping obtains an HC5 estimate within a range calculated over many resamples (typically 5000) drawn at random from the original sample, with replacement (Efron and Tibshirani, 1993). Despite the obvious advantages of the bootstrap method in not requiring a priori distribution assumptions, and the simplicity in calculating bootstrap confidence intervals, this is a data-demanding approach requiring at least 20 data points to define HC5 and associated confidence intervals (Grist et al., in press).

4.3. Bootstrap regression method

Bootstrap regression can be considered a compromise between the power of resampling and fitting an underlying distribution. This hybrid-technique allows the use of smaller toxicity datasets and the calculation of confidence intervals around the point estimate (Grist et al., in press). When there are few data, that do not adequately fit one of the standard distributions, bootstrap regression with an appropriate underlying model provides a suitable alternative.

4.4. Comparisons of the methods

Here we extracted 15 saltwater datasets (covering various modes of action) from the AQUIRE database. Multiple data for the same species were summarised as geometric means. The datasets were analyzed using the following methods:

1. log-normal model;
2. log-logistic model, L_{HC5} was calculated using the method of Aldenberg and Slob (1993);
3. standard non-parametric bootstrap method (Grist et al., in press); and
4. bootstrap regression based on a log-logistic regression model (Grist et al., in press).

We calculated HC5 values and lower confidence limits (L_{HC5}; 95% one-tailed) along with the appropriate regression parameters for each of the chemicals (Table 3). Results were compared visually by the goodness of 'curve fit' at the lower percentile region, and the parity and conservativeness of the HC5 and L_{HC5} values.

In general, the various methods yielded different HC5 and L_{HC5} values (Table 3). For example, the standard bootstrap was the best method for fitting a model distribution to the cadmium dataset (Fig. 3C) and it generated a substantially lower HC5 at 0.511 μg/l (Table 3). In contrast, the HC5 estimate for cadmium produced by the conventional log-logistic method was some 16 times higher than the value generated by the standard bootstrap (Table 3). Nevertheless, in most cases, differences in the HC5 and L_{HC5} values are within a factor of 2.

Bootstrap regression was found to be especially useful for small datasets where a good model fit is achieved. We combined a log-logistic regression with the bootstrap to improve the fit to the data. For example, both parametric models achieved a good fit to the small nickel dataset (Table 3; Fig. 4). Although the conventional log-logistic model implied a nickel HC5 estimate of 529.7 μg/l, bootstrap regression further improved the model fit but generated a more conservative HC5 at 321.7 μg/l (Table 3; Fig. 4).

The choice of an appropriate method for SSD analysis is important because the various methods may generate different HC5 and L_{HC5} values. In order to obtain the best HCp estimate, we suggest that the collected data should be analyzed by all four approaches. The method providing the best fit, especially in the lower tail can be identified by considering the coefficient of determination (r^2) in tandem with visual inspection. Subsequently, the final HCp can be estimated more

Table 3
Comparison of the HC5 value and its lower confidence limit (L_{HC5}) for 95% one tail (equivalently, 90% two-tailed) confidence intervals calculated by the different approaches[a]

Chemicals	Ammonia	Cadmium	Chlordane	Copper	Dieldrin	Endosulfan	Lead	Lindane	Malathion	Nickel	Pentacholrophenol	Phenol	Potassium dichromate	Trichloroethane	Toluene
Sample size	14	31	8	24	33	25	36	35	28	9	30	28	33	12	7
Log-normal															
a	1.162	0.656	1.177	1.490	0.893	0.572	0.977	0.661	0.716	0.962	1.399	1.356	1.541	2.424	1.121
b	−5.185	−2.022	−0.985	−3.477	−1.302	−0.346	−3.733	−1.262	−1.638	−3.936	−3.871	−6.167	−6.437	−11.71	−5.344
r^2	0.964	0.913	0.945	0.984	0.847	0.904	0.904	0.914	0.932	0.979	0.980	0.941	0.961	0.841	0.981
HC5	1110	3.76	0.275	16.96	0.412	0.005	137.1	0.263	0.979	241.3	38.99	2166	1288	14206	1994
L_{HC5}	529.2	0.608	0.084	12.06	0.068	0.001	32.39	0.044	0.222	109.1	26.21	1046	772.0	5897	961.3
Log-logistic															
α	4.460	3.083	0.837	2.333	1.458	0.604	3.820	1.909	2.287	4.093	2.767	4.549	4.178	4.831	4.767
β	0.402	0.737	0.367	0.332	0.525	0.830	0.490	0.738	0.679	0.465	0.358	0.360	0.324	0.177	0.386
r^2	0.979	0.940	0.947	0.971	0.913	0.897	0.901	0.914	0.955	0.970	0.973	0.977	0.942	0.771	0.887
HC5	1892	8.22	0.570	22.67	0.817	0.014	238.2	0.545	1.941	529.7	51.52	3083	1675	20417	4271
L_{HC5}	442.4	1.07	0.039	6.842	0.060	0.001	62.01	0.072	0.167	25.24	19.34	834.8	691.0	8110	180.2
Basic bootstrap															
HC5	ND	0.511	ND	25.10	2.795	0.049	314.1	0.178	2.780	ND	49.79	1495	1796	ND	ND
L_{HC5}	ND	0.200	ND	24.00	2.356	0.040	20.70	0.170	2.047	ND	37.00	510.0	1700	ND	ND
Bootstrap regression based on log-logistic model															
HC5	1819	12.96	0.480	13.07	0.863	0.008	795.9	0.546	1.448	321.7	30.14	2670	919.7	16375	1605
L_{HC5}	238.9	0.983	0.034	5.652	0.384	0.001	120.7	0.087	0.327	33.04	13.97	1261	520.8	8354	330.0

Notes: Log-normal model: $Y = aX + b$ where Y is percentile in probit scale, X is concentration in log scale, a is the slope and b is a y-intercept.
Log-logistic model: $Y = 1/[1 + \exp^{(-(X-\alpha)/\beta)}]$, where Y is percentage of affected species, X is concentration in log scale, α and β are constants.
[a] Both values are expressed as μg/l. Model parameters for log-normal and log-logistic models are also presented. ND: not defined, due to small sample size ($n < 20$).

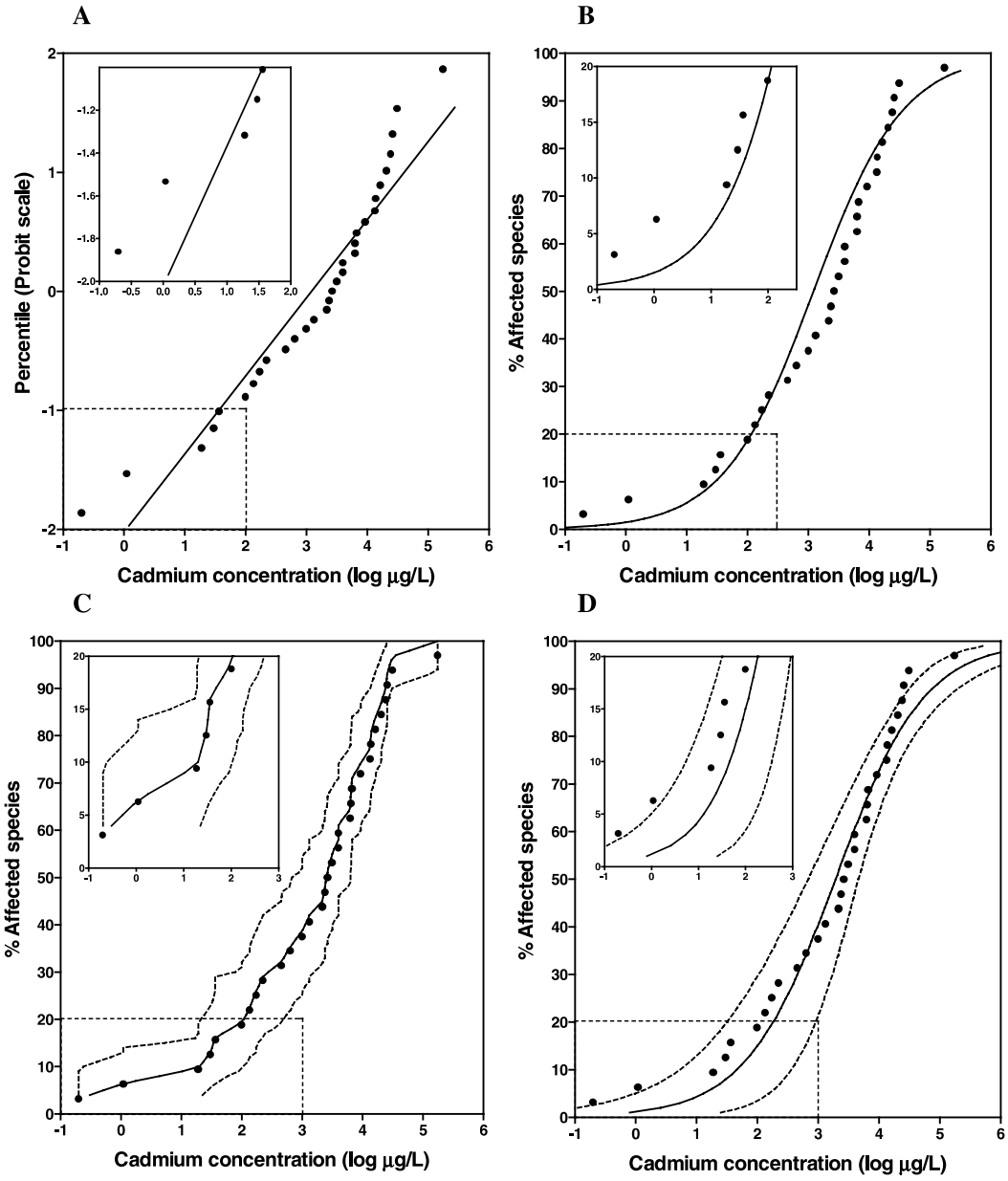

Fig. 3. SSDs for cadmium obtained by different methods: (A) log-normal (y-axis in probit scale); (B) log-logistic; (C), standard bootstrap and (D) bootstrap regression. Solid line denotes HCp while dashed lines are the two-tailed 90% confidence intervals. Inserted figures amplify the lower percentile region. Curves fitted by both the conventional model approaches deviated substantially from actual data points at the lower end of the SSD (A, B). In contrast, the HCp calculated by the standard bootstrap closely followed the entire data distribution (C). The bootstrap regression technique generated a curve which closely matched the conventional log-logistic model (D).

reliably by the best method. As a general guideline, (1) the standard bootstrap method is recommended when there are enough data (species) and (2) the bootstrap regression is especially useful for small datasets for which a good fit is achieved by a specified regression model.

5. Discussion

The diversity of SSD approaches developed is testament to the fact that toxicity data do not uniformly fit into a single class of model. This coupled with the variety of data summaries leaves a wide range of approaches available to the risk assessor. However, it appears that in the recent literature the full variety of methods are generally not applied. Table 4 indicates that the use of the most sensitive toxicity data fitted by the log-normal model is most commonly used. We have demonstrated that data quality, quantity, summary and model choice all influence analyses to a greater or lesser extent. We believe an acknowledgement of these differences is important to the transparency of

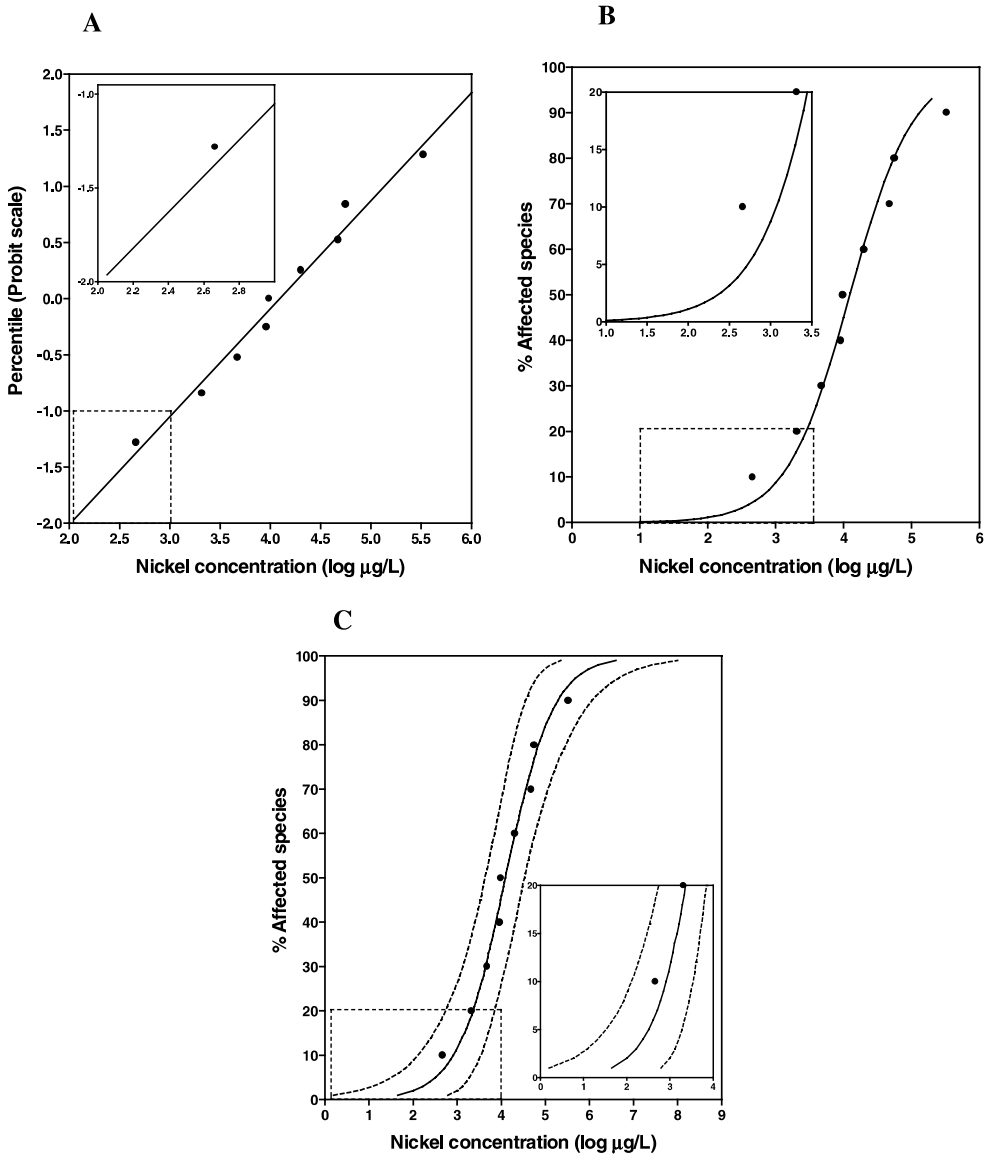

Fig. 4. SSDs for nickel using the different methods: (A) log-normal (y-axis in probit scale); (B) log-logistic and (C) bootstrap regression. Solid line denotes HCp while dashed lines are the two-tailed 90% confidence intervals. Inserted figures amplify the lower percentile region. For such a small dataset, a good fit was achieved by both parametric methods ($r^2 = 0.98$ [log-normal] and 0.97 [log-logistic]) (A, B). The bootstrap regression gave a better fitting curve (thus HCp estimate) in the lower tail compared to the log-logistic method (C).

Table 4
Published probabilistic risk assessments data summaries and distribution applied to SSD analyses

Substance	Media	Data summary	Model	Reference
Atrazine	Freshwater	Most sensitive	Log-normal	Solomon et al. (1996)
Diquatdibromide	Freshwater	Most sensitive	Log-normal	Campbell et al. (2000)
Phthalates	Freshwater, saltwater, soil	Geometric	Log-logistic	van Wezel et al. (2000)
Pyrethroids	Freshwater	Geometric mean	Log-normal	Solomon et al. (2001)
Tributyltin	Freshwater and saltwater	Most sensitive	Log-normal	Hall et al. (2000)

the risk assessment process based on probabilistic methods.

We have shown that SSD analysis outputs (both log-normal and log-logistic) appear to stabilise with 10–15 data points, suggesting 10 data points as a minimum data requirement to generate reliable estimates upon which regulatory decisions may be based. At present the minimum data requirements are generally below 10

where SSDs are used to derive water quality objectives (e.g., OECD recommends at least 5 data points). We have also demonstrated that the quality of the input data can have an effect on SSD outputs, highlighting the need for stringent quality criteria for inclusion of data. Quality schemes such as used here (Box 1) or those being developed for ecological (ecotoxicological) benchmarks (Durda and Preziosi, 2000) need to be applied to SSD analyses.

Data summaries inevitably lead to a loss of useful information from the available dataset. An awareness of variability in single species tests is required despite the need to estimate an average response for each species. We currently lack a mathematical solution that will allow inclusion of all data whilst retaining protection at the species level. Therefore we are left with a trade-off in favour of simplicity and ease of use, which in itself may be an argument for the application of a modest safety factor to the results of an SSD analysis.

A wide range of different models are available as risk assessment tools. Each method will be more appropriate under different circumstances. For instance if there are sufficient data and a log-normal model fits well (especially at the lower tail), it will adequately describe risk and communicate implications for management strategies. However, in some circumstances the log-logistic distribution will provide a superior fit to the data and yield more realistic, and therefore useful, results. If both of these parametric descriptors do not adequately describe the data (fail to fit the data at the lower percentile region) then an assessor may wish to turn to a bootstrapping technique where, if sufficient data are available ($n > 20$ for HC5 or $n > 10$ for HC10), a normal resampling method may be applied (Grist et al., in press). Alternatively, if data are limited, a bootstrap regression procedure will provide a point estimate (HC5) and an associated confidence interval that could not be determined using standard bootstrapping, due to insufficient sample size.

In general, we recommend the consideration of all the techniques available so that regulatory decisions are only based on information from the most appropriate method. Risk assessments often present problems that may be substance or site specific, so flexibility in procedures can be a positive advantage. We believe, however, that application of the range of methods needs standardisation to ultimately ensure the transparency of the entire process.

Acknowledgements

We are grateful to the CEFIC Long Range Initiative for Funding. E.P.M. Grist is supported by the Natural Environment Research Council, UK. K.M.Y. Leung thanks The Croucher Foundation, Hong Kong for providing a postdoctoral research fellowship.

References

Aldenberg, T., Slob, W., 1993. Confidence limits for hazardous concentrations based on log-logistically distributed NOEC toxicity data. Ecotoxicology and Environmental Safety 25, 48–63.

ANZECC and ARMCANZ, 2000. Australian and New Zealand guidelines for fresh and marine water quality. National Water Quality Management Strategy Paper No 4. ANZECC and ARMCANZ, Canberra.

Campbell, K.R., Bartell, S.M., Shaw, J.L., 2000. Characterizing aquatic ecological risks from pesticides using a diquat dibromide case study. II. Approaches using quotients and distributions. Environmental Toxicology and Chemistry 19 (3), 760–774.

Crane, M., Newman, M.C., 2000. What level of effect is a no observed effect? Environmental Toxicology and Chemistry 19 (2), 516–519.

Crane, M., Sorokin, N., Wheeler, J.R., Grosso, A., Whitehouse, P., Morritt, D., 2001. European approaches to coastal and estuarine risk assessment. In: Newman, M.C., Roberts Jr., M.H., Hale, R.C. (Eds.), Coastal and Estuarine Risk Assessment. Lewis Publishers, London, pp. 15–39.

Durda, J.L., Preziosi, D.V., 2000. Data quality evaluation of toxicological studies used to derive ecotoxicological benchmarks. Human and Ecological Risk Assessment 6 (5), 747–765.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall, New York, USA.

Feibicke, M., Ahlers, J., 2001. Environmental effects assessment for substances with large database some more detailed explanations. Prepared for EU-ECB special technical meeting: PNEC derivation for data-rich substances. Umwelt Bundes Amt, Berlin.

Forbes, T.L., Forbes, V.E., 1993. A critique of the use of distribution-based extrapolation models in ecotoxicology. Functional Ecology 7, 249–254.

Giesy, J.P., Solomon, K.R., Coates, J.R., Dixon, K.R., Giddings, J.M., Kenaga, E.E., 1999. Chlorpyrifos: ecological risk assessment in North American aquatic environments. Review of Environmental Contamination and Toxicology 160, 1–129.

Grist, E.P.M., Leung, K.M.Y., Wheeler, J.R., Crane, M., in press. Better bootstrap estimation of hazardous concentration thresholds to protect biological assemblages. Environmental Toxicology and Chemistry.

Hakanson, L., 1995. Optimal size of predictive models. Ecological Modelling 78, 195–204.

Hall, L.W., Scott, M.C., Killen, W.D., Unger, M.A., 2000. A probabilistic ecological risk assessment of tributyltin in surface waters of the Chesapeake Bay watershed. Human and Ecological Risk Assessment 6 (1), 141–179.

Jagoe, R.H., Newman, M.C., 1996. Bootstrap estimation of community NOEC values. Ecotoxicology 6, 293–306.

Leung, K.M.Y., Morritt, D., Wheeler, J.R., Whitehouse, P., Sorokin, N., Toy, R., Holt, M., Crane, M., 2001. Can saltwater toxicity be predicted from freshwater data? Marine Pollution Bulletin 42, 1007–1013.

Newman, M.C., Ownby, D.R., Mezin, L.C.A., Powell, D.C., Christensen, T.R.L., Lerberg, S.B., Anderson, B.-A., 2000. Applying species-sensitivity distributions in ecological risk assessment: assumptions of distribution type and sufficient numbers of species. Environmental Toxicology and Chemistry 19 (2), 508–515.

Parkhurst, D.F., 1998. Arithmetic versus geometric means for environmental concentration data. Environmental Science and Technology 32 (3), 92–99.

Shao, Q.X., 2000. Estimation for hazardous concentrations based on NOEC toxicity data: an alternative approach. Environmetrics 11 (5), 583–595.

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

Smothers, C.D., Sun, F., Dayton, A.D., 1999. Comparison of arithmetic and geometric means as measures of a central tendency in cattle nematode population. Veterinary Parasitology 81, 211–224.

Solbe, J.F., Buyle, B., Guhl, W., Hutchinson, T., Länge, R., Mark, U., Munk, R., Scholz, N., 1993. Developing hazard identification for the aquatic environment. Science of the Total Environment Supplement, 47–57.

Solbe, J.F., Buyle, B., Guhl, W., Hutchinson, T., Kloepper-Sams, P., Länge, R., Munk, R., Scholz, N., Bontinck, W., Niessen, H., 1998. Analysis of the ECETOC aquatic toxicity (EAT) database. I. General introduction. Chemosphere 36 (1), 99–113.

Solomon, K.R., Baker, D.B., Richards, R.P., Dixon, K.R., Klaine, S.J., La Point, T.W., Kendall, R.J., Weisskopf, L.W., Giddings, J.M., Giesy, J.P., Hall, L.W., Williams, W.M., 1996. Ecological risk assessment of Atrazine in North American surface waters. Environmental Toxicology and Chemistry 15 (1), 31–76.

Solomon, K.R., Giddings, J.M., Maund, S.J., 2001. Probalistic risk assessment of cotton pythrethoids. I. Distributional analyses of laboratory aquatic toxicity data. Environmental Toxicology and Chemistry 20 (3), 652–659.

Steen, R.J.C.A., Leonards, P.E.G., Brinkman, U.A.T., Barcelo, D., Tronczynski, J., Albanis, T.A., Cofino, W.P., 1999. Ecological risk assessment of agrochemicals in European estuaries. Environmental Toxicology and Chemistry 18 (7), 1574–1581.

Stephan, C., Mount, D., Hansen, D.J., Gentile, J.H., Chapman, G.A., Brungs, W.A., 1985. Guidelines for deriving numerical national water quality criteria for the protection of aquatic organisms and their uses. US EPA, Environmental Research Laboratory, Duluth, MN.

Streiner, D.L., 2000. Do you see what I mean? Indices of central tendency. Canadian Journal of Psychiatry 45 (9), 833–836.

Toll, J., Adams, B., Brix, K., Burger, M., Cardwell, R., DeForest, D., Tear, L., Cordoso, T., 2001. Proposed approach for deriving predicted no effect concentrations for substances protecting aquatic ecosystems. Prepared for EU-ECB special technical meeting: PNEC derivation for data-rich substances.

van Straalen, N.M., van Rijn, J.P., 1998. Ecotoxicological risk assessment of soil fauna recovery from pesticide application. Review of Environmental Contamination and Toxicology 154, 85–141.

van Wezel, A.P., van Vlaardingen, P., Posthumus, R., Crommentuijn, G.H., Sijm, D.T.H.M., 2000. Environmental risk limits for two phthalates, with special emphasis on endocrine disruptive properties. Ecotoxicology and Environmental Safety 46 (3), 305–321.

Vega, M.M., Urzelai, A., Angulo, E., 1999. Minimum data required for deriving soil quality criteria from invertebrate ecotoxicity experiments. Environmental Toxicology and Chemistry 18 (6), 1304–1310.

Wagner, C., Lokke, H., 1991. Estimation of ecotoxicological protection levels from NOEC toxicity data. Water Research 25, 1237–1242.