Estimating species sensitivity distributions with the aid of expert judgements

Anthony O'Hagan*, Mark Crane†, Eric Grist‡
and Paul Whitehouse§

*Department of Probability & Statistics, University of Sheffield, Sheffield, S3 7RH, UK.
†Crane Consultants, Chancel Cottage, 23 London Street, Faringdon, Oxfordshire, SN7 7AG, UK.
‡CSIRO Marine Research, GPO Box 1538, Hobart, Tasmania 7001, Australia.
§Environment Agency, Evenlode House, Howbery Park; Wallingford, Oxfordshire, OX10 8BD, UK.

May 6, 2005

Abstract

A species sensitivity distribution (SSD) is the probability distribution of some measure of toxicity to a certain chemical in a population of animal species. Given data consisting of estimates of toxicity for a number of species present in the habitat of interest, the SSD is typically estimated by assuming that these values are a random sample from a lognormal distribution and estimating the lognormal parameters. The principal deficiency of this approach is the assumption that the data are from a random sample of species. In practice, the species for which data are available are determined in non-random ways and are likely to be highly non-representative of the population.

We present a method of inference about SSDs that draws on expert judgement about which species are likely to be more sensitive to the chosen chemical. The expert judgements allow us to take some account of non-representativeness of the available data. We adopt a hierarchical random-effects model which recognises that species within the same family are likely to have similar sensitivities, and employ a Bayesian approach to analysis that allows direct inference about quantiles of the SSD.

Keywords: Bayesian inference; censored data; chlorpyrifos; environmental standards; HC5; hierarchical model; LC50; SSD; toxicology.

1 Introduction

Environmental quality standards are formulated by national and international agencies to protect animals and plants from damage by potentially toxic chemicals. The analysis developed here is in the context of water quality standards that seek to control the concentrations of chemicals in aquatic habitats. The objective of such a standard is to keep the concentration of a given chemical within a given habitat below a critical level, at which either none or only a small proportion of species living in that habitat will be harmed.

We consider the community of animal species present in a given aquatic habitat, and suppose that a suitable toxicological measure has been defined specifying the concentration of a given chemical at which each species would be harmed. Then the community of species is a population in the statistical sense, and the distribution of that measure in the population of species is its *species sensitivity distribution* (SSD) for the given habitat and chemical. A commonly used measure in toxicology is the LC50, which is the concentration at which 50% of individuals of a given species will be killed within a given time period. Although killing 50% of individuals is rather an extreme definition of 'harm', the greater availability of published LC50 values makes it possible to construct a useful SSD for this measure where it would not be practical for more meaningful measures of harm. For instance, in the example of Section 3 we use data on the LC50 for 96-hour exposure to the chemical chlorpyrifos.

Given a suitable SSD, a standard may be set to keep the concentration of the chemical below a level at which only a small proportion of species will be harmed. This level is the corresponding quantile of the SSD. For instance, the HC5 is defined as the 5-th percentile of the SSD, and therefore at least 95% of species will be protected if the concentration does not exceed the HC5. Estimation of the HC5 and other SSD quantiles is discussed by Wagner and Lokke (1991), Aldenberg and Slob (1993) and Aldenberg and Jaworska (2000).

The standard approach to estimating the SSD (see for example Newman et al 2000, Posthuma et al 2001, Grist et al 2002) assumes that the distribution is lognormal and that the available data are derived from a random sample of species from the statistical population. Then to estimate the SSD it is simply a matter of estimating the mean and variance of the normal distribution of log-concentrations. The two assumptions of lognormality and random sampling are both strong. In practice, it can often be argued that the SSD will be approximately lognormal, but there are rarely enough data points with which to test this assumption. This may lead to biased estimation of a relatively extreme quantile such as the HC5.

The assumption that data are from randomly sampled species is seriously inappropriate in practice (Forbes and Calow 2002, Maltby et al 2005). The species for which toxicity data are published will arise from various non-random mechanisms. First, some species are widely cultivated for laboratory use. These same species are tested with many different chemicals, which helps toxicologists to compare chemicals. They tend to be species that are known to be generally sensitive to water quality. For an investigator to test a species for which individ-

uals are not already available entails time-consuming collection from the wild, and relatively rare species are very unlikely to be tested. Second, species chosen for testing with a given chemical will tend to be ones that the investigators expect to be quite sensitive to that chemical.

We present a way to address this problem by use of expert judgements. We suppose that expert knowledge yields some measure of sensitivity for each species present in the habitat. We do not assume that this in any way produces even a perfect ranking of the species, only that there is some correlation between the expert measure and the true toxicological measure for the different species. By fitting a relationship between the expert measure and the toxicological measure, using the available data, we can predict the toxicological measure for all the species for which we do not have data. The stronger the correlation between the two measures, the less uncertainty there will be in these predictions, but even with a weak association the method can make some allowance for the non-representativeness of the available data. In Section 2 we present our model and analysis. We apply this to the 96-hour LC50 SSD for chlorpyrifos in three different habitats in Section 3. The results are discussed in Section 4.

2 Model and Bayesian analysis

Suppose that we have data comprising log-concentrations y_i, $i = 1, 2, \ldots, n$, for n species. We also have expert judgement in the form of a sensitivity measure. However, it is not practical to expect experts to specify a value for every one of the thousands of distinct species that may be found in a given habitat. Instead, we suppose that we have a value x_j, $j = 1, 2, \ldots, N$, for each of N groups, where in effect the expert is giving a representative sensitivity value to apply to all the species in the group. Such a group might be a genus, a family, a suborder or some other taxonomic collection. In general, we refer to such a group as a *taxon*; so we have expert assessments of sensitivity for each of N *taxa*. We let t_i, $i = 1, 2, \ldots, n$, denote the taxon to which the species with observed value y_i belongs.

2.1 Model

The expert assessments x_j could be derived in any way, and be on any appropriate scale, but we assume that they are monotonically related to the typical true toxicological log-concentration for species in taxon j, which we denote by μ_j. Formally, our model is hierarchical with the following components.

First, we let

$$y_i \sim N(\mu_{t_j}, \sigma^2) \ .$$

This assumption implies a lognormal distribution for the concentration (e.g. LC50) for species within a given taxon. This is weaker than the usual lognormality assumption for the whole SSD, since the expert opinion allows for quite flexible mixing of these taxa distributions. Notice that the variance σ^2 accounts for both measurement error in the data and for variability between species. The

usual approach to fitting a SSD ignores error in the data, and in effect we will do so here when we build the SSD from our model. It would be relatively simple to incorporate measurement error if we had reliable error assessments for the individual data (or repeat measurements for the same species). In practice, given good toxicological data its measurement error should be much smaller than the part of σ^2 that is due to within-taxon species variation. For simplicity, we do not separate measurement error and species variability here. Notice also that we assume a common σ^2 value for all taxa.

Second, we suppose a linear regression relationship holds between the taxa means and the expert assessments.

$$\mu_j \sim N(\alpha + \beta x_j, \tau^2) \ .$$

The variance τ^2 determines how accurately the expert assessments predict the true mean log-concentration for each taxon. Large τ^2 will mean that they have poor predictive value, and will increase the uncertainty in inferences about the SSD. It would be possible, given enough data, to estimate a non-linear relationship. It would be simple to extend the model to include other regression relationships but in practice this will rarely be useful; we retain the linear form here for simplicity.

Provided the t_is are not all distinct, so that there is some replication of taxa in the data, both σ^2 and τ^2 will be identifiable.

2.2 SSD and HC5

We now define the SSD in terms of this model. Just as it is not practical to suppose that experts can provide sensitivity assessments for all the individual species, it is not realistic to try to list all the species that might be present in the given habitat. Instead, we suppose that we have a list of all the taxa represented, and that these are the N taxa for which the experts have provided assessments. For taxon j, the proportion of species with true log-concentration values below some value y is $\Phi\left\{(y-\mu_j)/\sigma\right\}$. Now let w_j denote a weight to be attached to taxon j in constructing the SSD satisfying the condition $\sum_{j=1}^{N} w_j = 1$. In terms of the SSD, we treat w_j as representing the relative abundance of the different taxa, but it could also represent their relative importance for environmental protection. Then the SSD as a function of log-concentration has the cumulative distribution function

$$S(y) = \sum_{j=1}^{N} w_j \Phi\left\{(y-\mu_j)/\sigma\right\} \ . \tag{1}$$

The log-HC5 is the solution $\hat{y}_5$ to the equation

$$S(\hat{y}_5) = 0.05 \tag{2}$$

and the HC5 concentration is $\exp(\hat{y}_5)$. Any other desired quantile of the SSD can be defined analogously.

2.3 Inference

In principle it would be possible to analyse the data using classical, frequentist methods of inference, for instance by plugging maximum likelihood estimates of the μ_js and σ into (1) and then solving (2). However, because of the generally small sample size and the highly nonlinear nature of these equations, we cannot rely on the usual asymptotic theory to give a standard error or confidence interval around $S(y)$, $\hat{y}_5$ or $\exp(\hat{y}_5)$. Our preference is for a Bayesian analysis.

The model can readily be analysed using Markov chain Monte Carlo (MCMC) sampling, which is simple to implement in the WinBUGS package (Spiegelhalter et al, 2004). The MCMC method generates samples of $\left(\mu_1, \mu_2, \ldots, \mu_N, \sigma^2, \alpha, \beta, \tau^2\right)$ from the posterior distribution. Given a set of points $y_{(1)} < y_{(2)} < \ldots < y_{(m)}$, we can calculate $S(y_{(t)})$ for $t = 1, 2, \ldots, m$, for each MCMC sample, thus obtaining a sample from the posterior distribution of $S(y_{(t)})$. It is then simple to give a posterior credible interval for each $S(y_{(t)})$.

To obtain posterior inference about $\hat{y}_5$, it would be possible to use each MCMC sample to derive (1) and solve (2) numerically. This would give a sample from the posterior distribution of $\hat{y}_5$, from which an estimate and credible interval can be derived. However, a simpler approach uses the samples of $S(y_{(t)})$ values. Let $p_{(t)}$ be the proportion of MCMC sample values of $S(y_{(t)})$ that are less than 0.05. Then $p_{(t)}$ is the MCMC estimate of

$$P(S(y_{(t)}) \leq 0.05 \,|\, \mathbf{y}) = P(\hat{y}_5 \geq y_{(t)} \,|\, \mathbf{y}) \ .$$

Therefore the pairs $\left(y_{(t)}, 1 - p_{(t)}\right)$ estimate the posterior cdf of $\hat{y}_5$. It is simple to show that the sequence $1 - p_{(1)}, 1 - p_{(2)}, \ldots, 1 - p_{(m)}$ is increasing.

As always in a Bayesian analysis, it is necessary to formulate prior distributions for the parameters, in this case for α, β, σ^2 and τ^2. If there is substantive prior information about these parameters, it would of course be beneficial to use it, particularly in view of the limited data that will usually be available. Prior information may exist in practice about the magnitude of measurement error and species variability within taxa, to enable a proper prior distribution to be formulated for σ^2. We can generally say which sign β should have, since while it may be possible that the expert information is not well correlated with true toxicity measurements it should at least be true that higher expert assessments of sensitivity equate on average to higher true sensitivity. Where prior information is not thought to be sufficiently informative to affect the analysis, standard weak prior distributions may be used as in the following example.

3 Example

3.1 Data

A literature search found $n = 17$ measurements y_i of the 96-hour LC50 for the organophosphorous insecticide chlorpyrifos relating to freshwater aquatic species found in the UK. Seventeen biologists employed by the Environment

Agency or belonging to the Freshwater Biological Association were asked to score the sensitivity to chlorpyrifos of each of $N = 96$ taxonomic groups (covering most species likely to be found in UK freshwater habitats) on a scale from 1 (insensitive) to 8 (highly sensitive). A single weighted average score was derived from these to provide the expert assessments x_j. Lists of taxa were drawn up for three different habitats, a fast-flowing stream, a slow-flowing lowland river and a static pond or ditch. In each case, equal weights w_j were assigned. Details of all these data, including the elicitation of expert judgements and the averaging process, are given in Grist et al (2005).

Figure 1 plots the observations y_i against the corresponding expert assessments x_{t_i}. Points encircled are for different species in the same taxon. The line shows an empirical linear relationship, which has negative slope as expected but the expert judgements are generally not strongly correlated with the true toxicity. It is clear that the data do not support any more complex relationship being fitted. The replication of species within a taxon supports the assumption of constant σ^2 in the hierarchical model in which there is separate between-taxa and within-taxa variability.

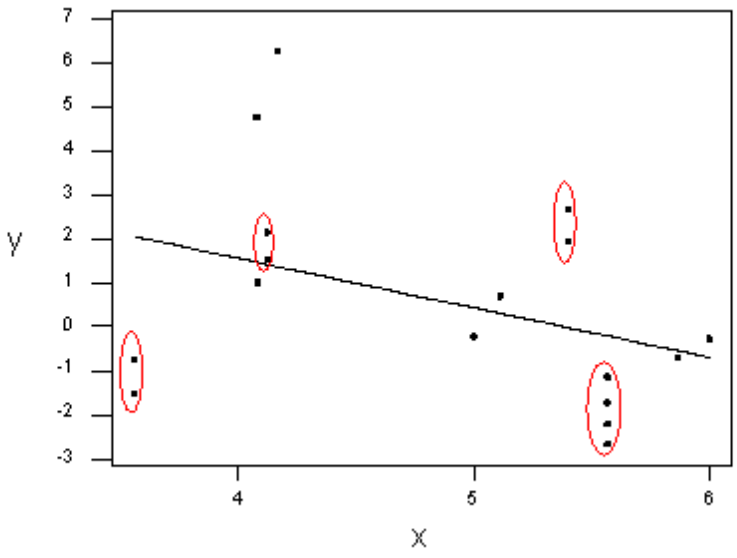

Figure 1. Toxicological data plotted against expert assessments

With regard to the representativeness of the sample data, we note that the experts assessed mean sensitivities of 5.33, 5.18 and 5.00 for the taxa present in each of the three habitats, whereas their mean assessed sensitivity for the 11 taxa in the data was 4.81. In general, then, in the experts' judgement the data under-represent the general sensitivity of the population.

3.2 Initial analysis

To use the standard method to estimate the SSD, we should treat the data as a random sample of species from the population. This is highly improbable, and particularly if we note that four of the data points are from the single taxon Gammaridae, which are shrimps. These are widely used in toxicological studies, but it is unreasonable to suppose that they make up about a quarter of all species in UK freshwater habitats. Actually, although all 17 species (from 11 taxa) are found in the slow-flowing river habitat, only 12 (from 7 taxa, including the Gammaridae) are appropriate to the fast-flowing stream and 11 (9 taxa) to the static pond. It is usual in eco-toxicology to ignore this distinction and to use all the available data to construct the SSD, which is then deemed applicable to a range of habitats. In our example, this is strictly applicable to the slow-flowing lowland river habitat.

We applied the usual approach, treating the 11 taxa as randomly sampled from the population, using sample means of the y_is for the replicated taxa. Fitting a normal SSD to these log-concentration summary data yields an estimated $N(1.153, 6.372)$ distribution, with log-HC5 at $1.153 - 1.645 \times \sqrt{6.372} = -2.999$. So the estimated HC5 concentration for the slow-flowing river is $\exp(-2.999) = 0.050$. Treating the 17 data points as a random sample would have given an estimated HC5 of 0.032, showing the sensitivity of the standard approach to the assumption of random sampling.

Using our model, we conducted a Bayesian analysis using weak prior information on all the parameters, except that we constrained β to be negative. The posterior distribution of β had expectation -1.5, with 95% probability between -3.3 and -0.1. Figure 2 shows the posterior mean and 95% credible intervals for the slow-flowing lowland river SSD, plotted against log-concentration, together with the $N(1.153, 6.372)$ distribution that was fitted to the taxa means.

Several features of Figure 2 are worth noting. First, there is considerable uncertainty about the SSD, shown by the wide credible bounds. Comparing the posterior mean with the conventionally fitted normal distribution based on the 11 taxa means, we see that in general the posterior mean is shifted to the left, and in particular it will give a lower estimated HC5. This is due to the fact that the expert assessments of sensitivity for the 11 taxa in the sample were on average rather lower than for the full set of taxa found in this habitat. The posterior mean SSD is also rather more spread out, reflecting the fact that the expert assessments are more spread for the full set of taxa than for the 11 in the sample.

The solid line in Figure 3 shows the posterior cumulative distribution function of the log-HC5. The posterior median HC5 is at $\exp(-3.84) = 0.021$, which is substantially lower than the 0.05 estimated by the fitted lognormal SSD. However, the distribution is clearly strongly skewed, and in order to have 90% posterior probability of protecting 95% of species, the concentration would need to be below $\exp(-6.33) = 0.002$. Hence the implications for environmental regulation of incorporating the expert knowledge in this analysis are demonstrably substantial.

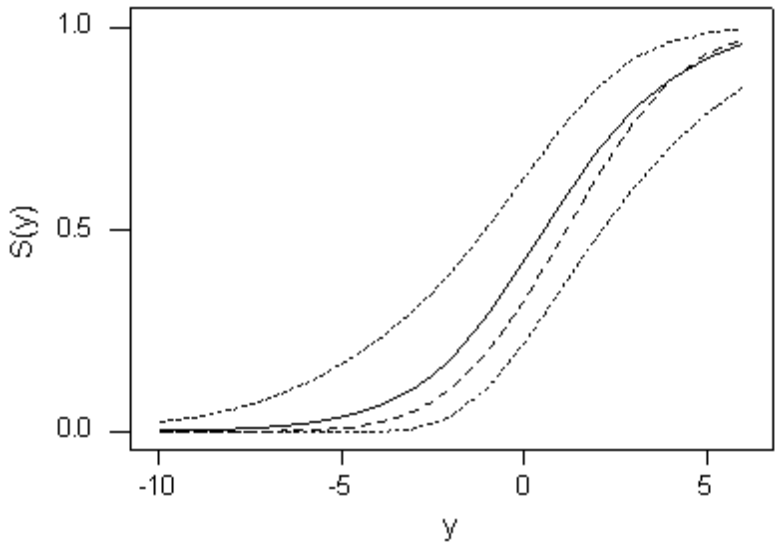

Figure 2. SSD for slow-flowing lowland river, based on 17 data points. Solid line, posterior mean; dotted lines, posterior 95% bounds; dashed line, fitted normal. Plotted against log-concentration.

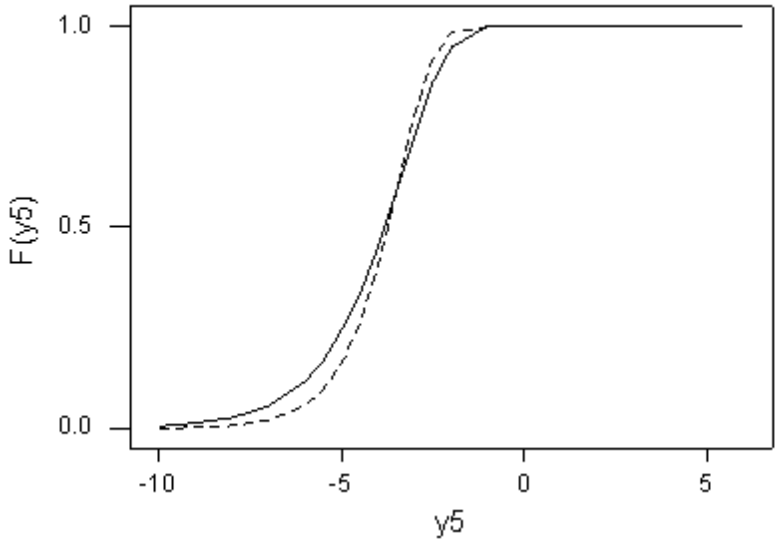

Figure 3. Posterior distribution of $\hat{y}_5$ for the slow-flowing lowland river. Solid line, based on 17 data points (Section 3.2); dashed line, based on 21 points (Section 3.3).

3.3 New data

In order to test the model and to improve estimation of the SSD, four additional species were tested in the laboratory. These were chosen in the light of the expert assessments (and reasonable availability), concentrating on species that the experts judged to be quite highly sensitive, including two species with higher assessed sensitivities than any in the original data. Three LC50 determinations provided new data points that could be plotted on Figure 1 at $(5.79, -3.35)$, $(6.04, -0.14)$ and $(6.27, -0.80)$. The last two points lie close to the fitted regression line in Figure 1. The first lies well below it, but is still comfortably within the posterior predictive 90% credible interval for this observation, based on the original 17.

The fourth new species produced an incomplete determination of its LC50, because the experiment did not include sufficiently high concentrations. The result for this species is simply that $y_{21} > \ln(10) = 2.30$. Its expert assessment value was 5.21, so this lies a little above the line in Figure 1, but again not sufficiently far from it to cast doubt on the model. It is not difficult to include this censored data item in the Bayesian analysis.

Repeating the analysis with these 4 new data points we find that the posterior mean of β is essentially unchanged, confirming again that the new data accord well with the model. Figure 4 shows the estimated SSDs on the log-concentration scale for the three habitats.

Credible bounds on these lines are a little narrower than in Figure 2, so it might appear that there is no proven difference between the habitats. However, the curves are not independent. At each point on the curve there is at least a 60% posterior probability that the SSD for the fast-flowing stream lies above that for the slow-flowing river, and at least a 70% probability that it lies above that for the static pond. So there is evidence (based in part on the expert assessments, and the fact that these have a demonstrated correlation with the true toxicological measurements) that the fast-flowing stream is the most sensitive of the three habitats.

The new posterior cumulative distribution function for the log-HC5 for the slow-flowing river is shown as a dashed line in Figure 3. There is now clearly less uncertainty about $\hat{y}_5$. The median HC5 value is $\exp(-3.75) = 0.024$, which is almost the same as with the original 17 data points, but the 10-th percentile doubles to $\exp(-5.5) = 0.004$.

For further details of this example, see Grist et al (2005).

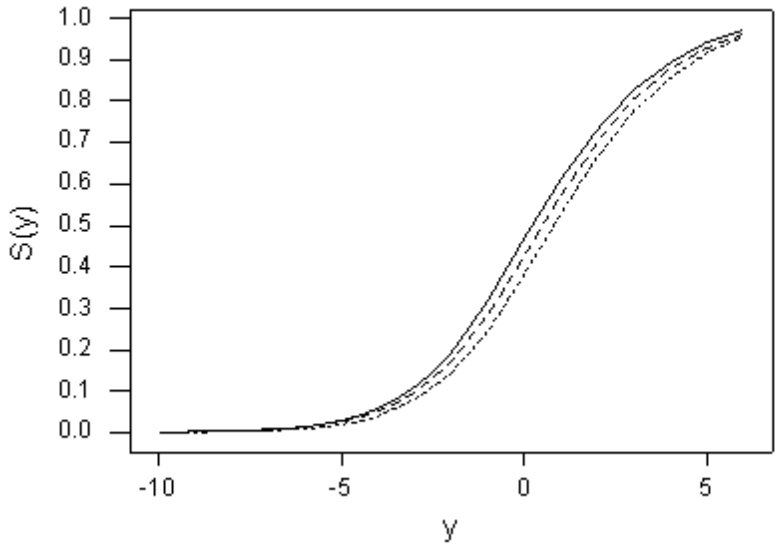

Figure 4. Posterior mean SSDs for three habitats, based on 21 data points. Solid line, fast-flowing stream; dashed line, slow-flowing lowland river; dotted line, static pond or ditch.

4 Discussion

We have presented a new method of estimating species sensitivity distributions that incorporates expert judgements of the relative sensitivities of different species. This approach is able to account for the non-random nature of the selection of species for which toxicological measurements are available. In the example which motivated this work, the expert assessments do not correlate very strongly with the toxicological data, but are nevertheless useful. There remains substantial uncertainty surrounding the estimated SSD, but a strength of the approach is that this uncertainty can be quantified clearly. Similarly, there is substantial uncertainty about low quantiles of the SSD, such as the HC5, that would be relevant to environmental regulation. Again, however, it is a strength of the method that this uncertainty is quantified clearly, without the use of asymptotic approximations, and potentially important skewness in the distribution of the HC5 is revealed.

In the example, extra data were gathered to test the model. The fit of these data to the model was good, and they confirmed the validity of the expert assessments. Despite the inherent variability in such data and the smallness of the original sample, the posterior inferences remained stable after adding the

new data. In contrast, the conventional method of fitting a lognormal SSD would have produced appreciably different answers before and after incorporating these data.

It is not straightforward to acquire the expert judgements that were used in this analysis, but the added value of these assessments is substantial in view of the very limited toxicological data that are generally available about any given chemical.

Acknowledgements

This research was funded by the Natural Environment Research Council under grant reference GST/02/2062 and the Environment Agency of England and Wales. We thank Environment Agency staff and members of the Freshwater Biological Association for help during the elicitation exercise.

References

Aldenberg T, Jaworska JS. (2000). Uncertainty of the hazardous concentration and fraction affected for normal species sensitivity distributions. *Ecotoxicology and Environmental Safety* **46**: 1–18.

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

Forbes VE, Calow P. (2002). Species sensitivity distributions revisited: A critical appraisal. *Human and Ecological Risk Assessment* 8: 473–492.

Grist EPM, Leung KMY, Wheeler JR, Crane M. (2002). Better bootstrap estimation of hazardous concentration thresholds for aquatic assemblages. *Environmental Toxicology and Chemistry* **21**: 1515–1524.

Grist EPM, O'Hagan A, Crane M, Sorokin N, Sims I, Whitehouse P. (2005). Comparison of frequentist and Bayesian freshwater species sensitivity distributions for chlorpyrifos using time-to-event analysis and expert elicitation. Submitted to *Environmental Science and Technology*.

Maltby L, Blake N, Brock TCM, Van den Brink PJ. (2005). Insecticide species sensitivity distributions: importance of test species selection and relevance to aquatic ecosystems. *Environmental Toxicology and Chemistry* **24**: 379–288.

Newman MC, Ownby DR, Mezin LCA, Powell DC, Christensen TL, Lerberg SB, Anderson BA. (2000). Applying species-sensitivity distributions in ecological risk assessment: Assumptions of distribution type and sufficient numbers of species. *Environmental Toxicology and Chemistry* **19**: 508–515.

Posthuma L, Suter GWH, Traas TP. (2001). *Species Sensitivity Distributions in Ecotoxicology*. Lewis: Boca Raton, FL, USA.

Spiegelhalter DJ, Thomas A, Best N, Lunn D. (2004). WinBUGS User Manual, version 1.4.1. http://www.mrc-bsu.cam.ac.uk/bugs/ and http://mathstat.helsinki.fi/openbugs/.

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