Species Non-Exchangeability in Probabilistic Ecotoxicological Risk Assessment

Peter S. Craig and Graeme L. Hickey,

Dept. of Mathematical Sciences, Durham University, Durham, United Kingdom.

Robert Luttik

The National Institute for Public Health and the Environment, Bilthoven, The Netherlands.

Andy Hart

The Food and Environment Research Agency, Sand Hutton, York, United Kingdom.

Summary. Current ecotoxicological risk assessment for chemical substances is based on the assumption that tolerances of all species in a specified ecological community are *a priori* exchangeable for each new substance. We demonstrate non-exchangeability using a large database of tolerances to pesticides for fish species and extend the standard statistical model for species tolerances to allow for the presence of a single species which is considered non-exchangeable with others. We show how to estimate parameters and adjust decision rules used in ecotoxicological risk management. Effects of parameter uncertainty are explored and our model is compared to a previously published less tractable alternative. We conclude that the model and decision rules proposed are pragmatic compromises between conflicting needs for more realistic modelling and for straightforwardly applicable decision rules.

Keywords: Exchangeability; Risk Assessment; Ecotoxicology; Species Sensitivity; Assessment Factors

1. Introduction

Much of modern statistics is concerned with models of increasing complexity, with goals of achieving greater realism and with addressing more complex inferences. However, some areas of risk management and decision making, such as ecotoxicological risk assessment (ERA), are resistant to such complexity and are unwilling to use rules which do not take simple intuitive forms. We examine ERA and show how a weakness in standard modelling can be addressed pragmatically, leading to adjustments to standard decision rules which should be comprehensible and usable by risk managers. Such procedures are more likely to be acceptable and therefore to be adopted.

ERA is an important tool for restricting the potential ecological damage from chemical substances, such as general chemicals or pesticides, while still permitting industry and agriculture to use them to their advantage. This has gained wider attention since the phased introduction of the new REACH regulation (EC, 2006) in 2007. It is required that manufacturers and importers gather information on the properties of all their substances, which will allow their safe usage. One such safety issue is the impact of environmental exposure to the substance, controlled or otherwise, on ecological (multi-species) communities, e.g. freshwater species. We defer a detailed discussion of ERA, and the underlying statistical model accepted by regulators, to Section 2 and proceed here with a simplified description of the statistical problem.

A simple view of the statistical aspects of ERA is that each substance defines a population of tolerances, expressed as concentrations or doses, where the tolerance is an attribute of a species rather than of individuals. We wish to determine a concentration or dose, known here as the environmental level of concern (ELC) for a substance, below which adverse effects are unlikely to occur to the ecological community being considered. However, practicalities and ethics mean that tolerances are measured for only a small number of species. A number of different approaches have been proposed for determining the ELC. The simplest is to divide the lowest measured tolerance by an assessment factor — an arbitrarily defined large fixed number which conservatively accounts for variability and uncertainty. This is motivated by the 'precautionary principle' which, in the context of ERA, Forbes and Calow (2002a) define as 'applying controls to chemicals in advance of scientific understanding if there is a presumption that harm will be caused'. A more refined approach, which we follow, is to adopt a simple statistical model for the measured tolerances which are treated as a random sample from a population of species tolerances and to use the model to help determine the ELC.

In practice, the species measured are not chosen randomly but the same procedure is followed, based effectively on the more realistic assumption, familiar to the Bayesian community, that all species tolerances for the new substance are *a priori* exchangeable. However, there is a body of informal evidence that the assumption of exchangeability is invalid, particularly in relation to pesticide exposure for one fish species, *Oncorhynchus mykiss* (rainbow trout). We explore a sequence of issues necessary to gaining a good view on how practically to allow for non-exchangeability in ERA: testing for non-exchangeability, tractable extension of standard modelling, estimation of hyper-parameters representing non-exchangeability and variance heterogeneity, risk measures and rules for determining the ELC, defensibility of a key assumption and alternative models for non-exchangeability.

The crux of the issue is that simplicity may be better than complexity, even when simplicity results in some relative weaknesses. The take-up of more complex statistical methodology in ecotoxicology is slow. Moreover, the regulatory process is controlled indirectly by legislation and directly by the risk managers who are not research scientists but who are required to be able to defend the risk management process when it is scrutinised by commercial or consumer interests. Procedures which involve relatively small adaptations of familiar techniques are seen to be more transparent and to be more defensible. Thus our focus is on the detection of non-exchangeability and on *tractable* ways to adapt current ERA methodology to allow for non-exchangeability in a *pragmatic* and *parsimonious* manner.

2. Ecotoxicological risk assessment

The decision making process in ERA is based on so-called risk characterisation (ECHA, 2008b) which involves: (i) estimation of the predicted exposure concentration (PEC) which might be found in an ecosystem, i.e. the wider interaction of the different ecological communities (assemblages of multi-species populations) and physical components (e.g. air and water) of an environment, for example a ditch; (ii) assessment of the degree to which the PEC may have adverse consequences on the communities.

Under current EU regulatory technical guidance, this fundamental approach to conducting ERAs for general chemicals (ECHA, 2008b) and pesticides (EC, 2002), which we denote generically as substances from here onwards, is based on a tiered process. At the lowest tier, the assessment is intended to be simple and economical, yet at the same time

robustly conservative. A high tier risk assessment, which is typically much more expensive, is triggered by the failure of lower tiers and generally calls for a detailed joint-probabilistic assessment of (i) and (ii) specific to each exposure scenario and ecological assemblage; the resulting ERA dossier is subsequently assessed carefully by expert scientists. Since it is not logistically practical to assess the risk to every species within every ecosystem, lower and intermediate quantitative tiers focus on the consequences for individual species based on a small number of tolerance measurements; the calculations act as a proxy for all ecosystems.

We focus on the intermediate tier of risk assessment. Here, the fundamental decision making criterion is: if the ELC $>$ PEC, the risk is deemed acceptable, otherwise permission for use is prohibited pending a higher tier assessment. We shall limit our discussion to aquatic ERA in order to simplify the language, but the methods discussed are applicable in a wider context, for example to bird-only risk assessment. In this section we provide details on two features of this problem: assessment factors and species sensitivity distributions, and elaborate further on the motivation for this research, non-exchangeability.

2.1. Assessment factors

Exposure is expressed as a concentration of the substance in water, and toxicity of the substance to a specific species (or genus) type is described in terms of a 'tolerance' concentration which yields a specific effect. A common choice is the median effect concentration (EC_{50}). This is the concentration which is statistically estimated to *affect* 50% of individuals for a single-species population in some fixed time period (often 24–96 hours) with respect to some chosen relevant measurable ecological endpoint, such as mortality. Species tolerance values for a specific substance, collectively referred to as *toxicity data*, will usually be estimated, and subsequently treated as known, only for a very small number of distinct species.

The standard first tier deterministic procedure determines the ELC by dividing the lowest measured tolerance by an 'assessment factor'. This is a positive fixed number (usually a power of 10 such as 1000) defined in the appropriate regulatory technical guidance document and which is intended to allow for: (i) variation between and within species; (ii) differences between acute and chronic sensitivity; and (iii) extrapolation from laboratory (i.e. single species tolerance) to field (i.e. ecosystems) impact. However, little or no justification is provided for its magnitude, leading to ambiguity about the actual level of intended protection (Forbes and Calow, 2002a).

2.2. Species sensitivity distributions

Considerable attention has been given to probabilistic techniques in order to derive ELCs. The fundamental underlying concept is the 'species sensitivity distribution' (SSD; Posthuma et al. 2002), which, for a specific substance, is a distribution modelling the interspecies variability of tolerance in an ecological community, thus providing a way, separate from any use of assessment factors for other purposes, to formally relate the tolerances of tested species to those of other untested species. There is no consensus on how to define the ecological community; Aldenberg et al. (2002) call this 'the Achilles heel of the *SSDeology*'. A weakness of the concept is the failure of measured species to represent communities (Forbes and Calow, 2002b), yet more refined approaches are stifled by limitations on data. Specific models which do address this, for example by weightings (Grist et al., 2006), are too complex for regular application in the intermediate tier of risk assessment. Consequently,

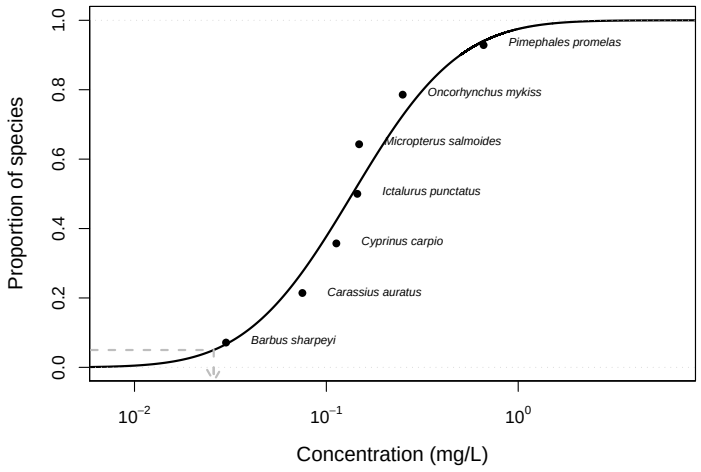

Fig. 1. Estimated SSDs for fish exposed to the herbicide *trifuralin*. Each point represents an EC_{50} value for the labelled species. The grey arrow indicates an estimate of the HC_5.

tolerance measurements for standard test species often act as proxies for many communities. It is the role of higher tier ERA to assess risk to (exposure site-) specific communities.

Standard parametric models for the SSD, motivated by pragmatism, are the log-normal distribution (Wagner and Løkke, 1991) and the log-logistic distribution (Aldenberg and Slob, 1993). Considerable attention (Hickey et al. 2008, 2009 and references therein) has been given to the problem of quantitative assessment of uncertainty concerning the p-th percentile of the SSD (denoted the HC_p). This is interpreted as the concentration which is hazardous to $p\%$ of species in an ecological community (Alexander and Fairbridge, 1999, p. 235), and for all intents and purposes defines the ELC subject to an additional SSD-specific assessment factor. A widely accepted protection goal is $p = 5$ (ECHA, 2008a). In Figure 1 we show an SSD estimated from tolerances for fish species exposed to the herbicide *trifuralin*.

The distributional assumptions and standard approaches to quantifying risk lead to rules for determining the ELC which typically all have the same form: the geometric mean of the toxicity data divided by a 'variable assessment factor' which is determined by the standard deviation of the SSD and the level of uncertainty. Determining this variable assessment factor has been the focus of recent research (Aldenberg et al., 2002; Hickey et al., 2009).

2.3. Non-exchangeability

The concept of SSDs involves many assumptions, some of which are un-testable (Forbes and Calow, 2002b). However, with a few exceptions such as Duboudin et al. (2004), one notable implicit assumption in the modelling literature is that, prior to observing the toxicity data for a substance, the tolerances of all species present in the ecological community are exchangeable. A direct implication of this is that information about relative rankings of species' tolerances in SSDs for other substances is uninformative about their relative rankings for the substance being assessed. An important statistical consequence of this is

that any measurements to be made for the substance may be considered to be a random sample from its uncertain SSD regardless of which species are to be measured.

The informal body of evidence (e.g. Dwyer et al. 2005) which suggests *O. mykiss* , and possibly other species, are non-exchangeable with respect to other fish species is supported by a recent report of the European Food Safety Authority (EFSA, 2005). Despite this, *O. mykiss* is a standard test species (Rand, 1995, p. 78).

The issue of (non)-exchangeability has largely been ignored in ERAs. Raimondo et al. (2008) issue caution about conducting ERAs based on the use of certain groups of species as proxies for all fish due to an apparent demonstration of higher tolerance. Stephan (2002) reports that one might purposefully populate estimated SSDs with recognisably less tolerant species to ensure conservatism, acknowledging that this *ad hoc* method violates SSD assumptions. Alternative methods such as bootstrapping described by Newman et al. (2000) may account for these effects, although it is not explicitly clear how. Grist et al. (2006) proposed the construction of community level SSDs as mixtures of distributions for taxonomic sub-groups, thereby acknowledging different tolerances of specific species groups.

A natural response of a statistical modeller (including some reviewers of this article) would be to abandon exchangeability and use a crossed random effects model (Goldstein, 1995, Chapter 8) incorporating both species and substance effects, although some adaptation of the standard model would be required to allow for observed heterogeneity in tolerance variability between substances. While that might succeed from a modelling perspective, it would substantially complicate the risk assessment procedure for several reasons. First, the incomplete factorial nature of any available database of measured tolerances would lead to highly confounded estimates of individual species and substance effects. Consequently, uncertainty attached to those estimates would be substantial and strongly correlated and would require careful propagation into decision rules. Secondly, it would not be possible to summarise the relevant information in an entire toxicity database through a small number of estimated parameters. The database would have to be made available to all participants in ERA and access to proprietary data would be an issue. Finally, the whole concept of the SSD and its use in ERA would require substantial reconsideration by ecotoxicologists. For example, unlike the current situation, making inferences about a percentile would require knowledge of the currently unspecified number of species in the ecological community. Overall, persuading risk managers to accept any resulting procedures would be extremely difficult.

3. Testing the assumption of exchangeability

EFSA (2005) provided an informal demonstration that *O. mykiss* may be non-exchangeable, showing graphically that its tolerance tended to be less than the geometric mean tolerance of other species measured on the same pesticide. We provide a more formal approach.

We investigate the null hypothesis that species tolerances are *a priori* exchangeable for each new substance, particularly pesticides. We propose two non-parametric tests, based on the ranks of an available toxicity database described below, motivated by the familiar sign and rank-sum tests for differences between two populations; the latter is more powerful but less robust as it is more sensitive to outcomes for individual substances. We chose a non-parametric approach to testing, despite the fact that the modelling approach in later sections is parametric, so that we could be sure that any test we used was actually providing evidence of non-exchangeability rather than evidence against parametric assumptions.

3.1. Data

The data we use were kindly supplied by The Dutch National Institute for Public Health and the Environment (RIVM) and comprise 1903 EC_{50} tolerance measurements for 172 distinct fish species and 379 different substances, in this case pesticides. The data, previously used by EFSA (2005), are a subset of a research database developed by De Zwart (2002) which has been amalgamated from many sources.

Henceforth, y_{ij} is the logarithm (base 10) of the tolerance of species j for substance i and the term SSD refers to the distribution of y_{ij} for fixed i. The number of species tested on substance i in the database is denoted n_i, and m_j is the number of substances on which species j has been tested. We also denote r_{ij} to be the rank of the measurement for species j amongst those tested on substance i, ties being assigned the average of the corresponding ranks. We use log-transformed tolerance for several reasons: (i) variability is stabilised (leading to additive errors); (ii) resulting distributions are often quite close to normal; and (iii) it is conventional in many areas of toxicology.

The data are by no means a complete factorial design; the EC_{50} has only been measured for 1903 of the possible 65,188 substance-species pairs. There are 143 substances for which $n_i = 2$, another 135 with $n_i \leq 5$, 64 with $6 \leq n_i \leq 10$, 30 with $11 \leq n_i \leq 20$ and 7 with n_i ranging from 21 to 47. From the species viewpoint, there are 74 for which $m_j = 1$, 22 with $m_j = 2$, another 26 with $m_j \leq 5$, 19 with $6 \leq m_j \leq 10$, 13 with $11 \leq m_j \leq 20$, 11 with $21 \leq m_j \leq 50$ and 7 individual species where m_j is respectively 54, 59, 76, 153, 160, 166 and 344. The last of these is *O. mykiss* which is the focus of much of this article.

3.2. Sign test

Under the null hypothesis of exchangeability, the tolerance of a species should be equally likely to appear above or below the median of the data for each substance. For each species, we can apply the binomial distribution to determine whether it occurs too often on one or other side. We ignore those substances where tolerance of the species equals the median; although this may reduce power, it leads to a simple exact conditional test.

For a species, calculate m^+ and m^- which are the numbers of substances for which the species tolerance respectively exceeds or is exceeded by the median of measured tolerances for the substance. Under the null hypothesis, conditional on the number of trials $m^+ + m^-$, m^+ has a binomial distribution with success probability $\frac{1}{2}$. We compute the two-tailed probability of obtaining a value as extreme as the observed m^+.

Results from applying this test to the RIVM database are displayed for the ten species with the smallest P-values in Table 1. One should be careful when interpreting the table. There is strong evidence against exchangeability but it does not guarantee that *O. mykiss* is the only such species presenting such a feature nor that it is the most, for want of a better word, biased species although it does identify it as a candidate. Clearly, there is more power to detect non-exchangeability when m is large but there are also species in the table which have not been tested very often. Note that, even if we apply the highly conservative Bonferroni correction to adjust the minimum P-value for multiple testing, the result is $172 \times 3.9 \times 10^{-15} = 6.7 \times 10^{-13}$.

3.3. Rank-sum test

As in the standard situation of comparing two populations, the rank sum test proposed here should be more powerful than the sign test. For species j, define the test statistic to

Table 1. Species with the smallest P-values for the sign test. m is the number of substances tested for the species, m^+ and m^- are the numbers of substances where the tolerance for the species respectively exceeds or is exceeded by the median.

Species	m	$m^+ + m^-$	m^+	$m^+/(m^+ + m^-)$	P-value
Oncorhynchus mykiss	344	301	83	0·28	3.9×10^{-15}
Carassius auratus	76	69	56	0·81	1.7×10^{-7}
Cyprinus carprio	166	150	103	0·69	5.6×10^{-6}
Heteropneustes fossilis	36	36	31	0·86	1.3×10^{-5}
Oncorhynchus clarki	42	41	10	0·24	1.5×10^{-3}
Pimephales promelas	160	147	93	0·63	1.6×10^{-3}
Carassius carassius	25	23	19	0·83	2.6×10^{-3}
Channa punctatus	17	16	14	0·88	4.2×10^{-3}
Clarias batrachus	17	16	14	0·88	4.2×10^{-3}
Salvelinus namaycush	35	33	8	0·24	4.6×10^{-3}

Table 2. Species with the smallest P-values for the rank sum test.

Species	m	P-value	Effect size
Oncorhynchus mykiss	344	8.6×10^{-12}	−0·42
Heteropneustes fossilis	36	1.9×10^{-7}	0·83
Carassius auratus	76	3.1×10^{-5}	0·68
Salvelinus fontinalis	33	1.3×10^{-4}	−0·58
Carassius carassius	25	1.6×10^{-4}	0·85
Oncorhynchus clarki	42	3.6×10^{-4}	−0·61
Clarias batrachus	17	4.0×10^{-4}	0·91
Salvelinus namaycush	35	2.4×10^{-3}	−0·59
Channa striata	10	3.9×10^{-3}	0·73
Perca flavescens	29	6.5×10^{-3}	−0·38

be the sum of r_{ij} over those substances for which the species has been tested. In effect, this gives more weight to substances for which more species have been tested. Conditional on n_i, under the null hypothesis, each r_{ij} is uniformly distributed on the integers 1 to n_i, provided there are no ties, and is independent for different values of i.

The exact null sampling distribution of the test statistic is computationally intractable but is easily approximated, either by Monte Carlo or a central limit theorem based normal approximation using the theoretical mean and variance which are easily obtained under the null hypothesis in the absence of ties. The difficulty with the former is that many of our P-values are very small and would require very many Monte Carlo repetitions. However, this is likely to happen only when m_j is large when we would expect the normal approximation to be more effective. As our activity is largely exploratory, we simply show P-values from the normal approximation in Table 2 for the RIVM database. Monte Carlo simulation with 10,000 repetitions did not give significantly different P-values; therefore, we did not attempt to adjust the normal approximation for ties. Also shown is an effect size for each species obtained by standardising each r_{ij} using the mean and standard deviation of the null discrete uniform distribution and computing the average value for each species. It provides some information about the average position of a species across a population of substances.

Interpretation of Table 2 is subject to the same caveat as for Table 1. It should be seen as providing further evidence of the apparent non-exchangeability of *O. mykiss* tolerances. Many of the same species appear and for those species the effect sizes in Table 2 are consistent with the relative sizes of m^+ and m^- in Table 1. The appearance of other species

indicates that the two tests emphasise different aspects of departures from exchangeability.

3.4. Focusing on O. mykiss

It is quite plausible that the exchangeability assumption is untenable from the perspective of statistical modelling and that all species are in fact non-exchangeable; if one eliminates all the *O. mykiss* data from the database one still finds clear evidence of non-exchangeability for the remaining species, based on both tests.

Instead we concentrate on the case of a single non-exchangeable species because our goal is tractable and useful decision rules rather than better statistical modelling. We consider the possibility of allowing for multiple non-exchangeable species in our final discussion. Our choice of *O. mykiss* as the single non-exchangeable species is justified by its special role in current regulation. It is a standard test species and therefore has greater potential than most species to influence risk assessment outcomes.

Aldenberg et al. (2002) showed that the rate at which the ELC changes as we perturb a single log-tolerance is greater for those log-tolerances which are less than the sample mean than for those which are greater. Therefore, non-exchangeability of *O. mykiss* deserves more attention than, for example, non-exchangeability of *Carassius auratus* (the goldfish), which is shown by Tables 1 and 2 to have a tendency to be less sensitive on average.

4. Modelling

We now suppose that there is a single special species which has non-exchangeable tolerance values. We revise our notation so that $y_i^\dagger$ denotes the log-tolerance of the special species for substance i and y_{ij} the log-tolerance for the other species.

Under *a priori* exchangeability, the standard model is that y_{ij} are independently sampled from $N(\mu_i, \sigma_i^2)$. We alter this only for the special species for which we specify $y_i^\dagger \sim N(\mu_i - k, [\phi\sigma_i]^2)$. Here k and ϕ are respectively location and scale adjustments and may be interpreted as specifying the predictive distribution for $y^\dagger$ were μ and σ to be known for a substance. They apply to multiple substances as only by so doing can we give them identifiable meaning; to be precise, k and ϕ^2 are respectively the averages across substances of $\mu - y^\dagger$ and $(y^\dagger + k - \mu)^2/\sigma^2$. Of course, there may be scientific grounds to have groups of non-exchangeability parameters for different classes of chemical, for example by the mode of action, but no available data supports this at present.

Our model for non-exchangeability derives from a different model proposed by EFSA (2005), for which the expected value of $y_i^\dagger$ was $\mu_i - k'\sigma_i$. In that model, scaling the offset k' of the mean by the standard deviation means that the expected percentile of the special species in the SSD is unaffected by variability of the standard deviation between substances. The EFSA (2005) model may be intuitively more appealing but we are not aware of any argument of principle favouring it. Moreover, unlike that model, our model leads later to tractable decision rules which are a key goal in this work. In Section 8, we assess whether the data favour one model over the other.

Obtaining values (or distributions) for k and ϕ requires the use at some stage of a database such as that provided by RIVM or of expert judgements. There is not uniform agreement about the role of such databases in risk assessment. It is clear that their use is acceptable for some purposes, such as the detection of non-exchangeability and therefore for estimation of k and ϕ, but some consider other uses to be unacceptable, for example construction of prior distributions for μ and σ by considering them to be drawn, along with

μ_i and σ_i, from hyper-populations of means and standard deviations. The lack of agreement in this area means that we consider two behavioural models in what follows:

M1 μ and σ unknown and varying between substances; database not used to provide prior information about μ and σ. See, for example, Aldenberg and Jaworska (2000).

M2 μ and σ unknown and varying between substances; σ assumed sampled from an inverse-gamma distribution with hyper-parameters α (shape) and β (rate); database for relevant other substances available to provide information about α and β; database not used to provide prior information about μ. See EFSA (2005).

M1 and M2 are not the only proposals in the literature. Aldenberg and Luttik (2002) suppose that μ varies but that σ does not and suggest determining a precise value for σ from expert opinion or a suitable database. EFSA (2005) consider consequences of uncertainty in estimating σ. However, there seems to be little justification for the assumption that σ does not vary, even for narrow definitions of chemical classes.

Under M1, each risk assessment is independent of others (apart from the sharing of evidence concerning the non-exchangeability parameters). This satisfies those who are wary of using evidence from previous assessments to form prior judgements. However, the small amount of data available for a typical risk assessment means that there will often be considerable benefit in exploiting previous experience to stabilise the estimate of σ for the current substance by incorporating the evidence about variation in values of σ from a database. No hyper-population of means is proposed in M2 as we have found the user-community to be resistant to the idea. Moreover, there is less to be gained than for the standard deviations as the RIVM database shows that variation in μ is high relative to typical values of σ, so that any proper prior for μ would typically be diffuse relative to the likelihood.

5. Hyper-parameter estimation

There are two groups of hyper-parameters: the non-exchangeability parameters k and ϕ which appear in both M1 and M2 and the heterogeneity parameters α and β which apply only to M2. In both cases, we use θ as a short-hand for the hyper-parameters.

We distinguish two groups of substances for which data may exist although they may not necessarily be publicly accessible. $\mathcal{G}_1$ is the group of substances, deemed to be relevant to the new substance, for which the tolerance of the special species has been measured. Under M2, we also need the collection $\mathcal{G}_2$ of substances considered relevant for estimating α and β. Note that under M2, we have to simultaneously estimate the non-exchangeability and heterogeneity parameters as they are linked through the likelihood. We shall assume that $\mathcal{G}_1$ is a subset of $\mathcal{G}_2$; although possible, it seems unlikely that substances would be considered relevant for estimation of non-exchangeability parameters but not for heterogeneity parameters. This assumption also simplifies the specification of prior distributions. In our example, as in EFSA (2005), we take $\mathcal{G}_2$ to be the complete collection of substances in the RIVM fish database and $\mathcal{G}_1$ to be the subset of all those where tolerances were measured for *O. mykiss* and at least 2 other species. This restriction, which was applied for direct comparability with a frequentist estimation approach in EFSA (2005), is not strictly necessary but provides more reliable information about the parameters.

In principle, under either behavioural model, one might elicit proper prior distributions for the hyper-parameters from a risk manager but this is unlikely in practice as aside from lack of time and expertise, it could constitute a conflict of interest and the risk manager

Table 3. MAP estimates for hyper-parameters k, ϕ, α and β with posterior standard deviations in parentheses.

	k	ϕ	α	β
M1	0.195 (0.019)	0.702 (0.073)	—	—
M2	0.205 (0.030)	0.656 (0.066)	1.52 (0.24)	0.315 (0.076)

would potentially be exposed to pressure from vested interests. In any case, we expect there to be significant amounts of data in both $\mathcal{G}_1$ and $\mathcal{G}_2$, and so we do not expect inferences to be very sensitive to the choice of prior distributions for the hyper-parameters. Under M1, we use independent improper prior distributions $\pi(k, \phi) \propto 1$ and $\pi(\mu_i, \sigma_i^2) \propto \sigma_i^{-2}$ for $i \in \mathcal{G}_1$. The latter is seen by many as the practical version of the Jeffreys prior and has been used in other Bayesian SSD literature, e.g. Aldenberg and Jaworska (2000) and EFSA (2005), where, as a consequence, frequentist and Bayesian risk calculations coincided. Under M2, the distribution of σ_i is determined by α and β and we again take $p(\mu_i) \propto 1$. For the heterogeneity hyper-parameters, we take $p(\alpha, \beta) \propto 1$ for $\alpha > 0, \beta > 0$.

With these prior specifications, substances are conditionally independent given the hyper-parameters and so their joint posterior distribution is a sufficient summary of the database when considering a new substance. This sufficiency means that the posterior distributions can be published and used without requiring open access to the databases from which they are derived (as was the case in EFSA 2005). In principle the posterior distributions should be updated whenever more data becomes available, for example every time a new substance is assessed. In practice, however, the same distributions will be used for many risk assessments for several reasons: (i) unavailability of raw data for re-estimation on the fly; (ii) infeasibility of sharing all data to ensure that everyone makes the same updates; (iii) lack of resources to re-appraise values.

Under both M1 and M2, the prior distribution and likelihood are now fully defined but we need to integrate out the nuisance parameters $\{\mu_i, \sigma_i^2\}$ to obtain the un-normalised marginal posterior density of the hyper-parameters. The posterior densities are briefly derived in Appendix A.1 and may be maximised numerically to obtain MAP (maximum a posteriori) estimates and the corresponding Hessian matrix.

Estimates and approximate posterior standard deviations are shown in Table 3. Values of k and ϕ are similar for M1 and M2, suggesting that information about non-exchangeability is largely uninfluenced by the introduction of a model for variance heterogeneity. Uncertainties attached to the estimates do not seem large; consequences for determination of ELC values are considered more formally in Section 7. The positive estimate of the offset hyper-parameter k suggests that *O. mykiss* tends to be a sensitive species having tolerance below the median of the SSD. Interpretation of ϕ is more difficult; however, $\phi < 1$ suggests that the SSD percentile for *O. mykiss* is less variable than for other species and leads to increased weight for the corrected tolerance in estimating the mean of the SSD. Overall, the estimates are consistent with previous informal suggestions that *O. mykiss* tends to be sensitive.

Our somewhat arbitrary choice of prior distribution for the hyper-parameters led us to investigate sensitivity to that choice by trying other prior distributions. For k we tried $p(k) \propto 1/(0.01 + k^2)$ which strongly favours values of k near 0 and $p(k) \propto (0.01 + k^2)$ which strongly favours large values of k. Similarly, for the other components of θ, which are all positive, we tried $p(\theta_i) \propto \theta_i$ and $p(\theta_i) \propto 1/\theta_i$. There were 4 alternative prior distributions for M1 and 16 for M2. In all cases the MAP estimates differed from those in Table 3 by

less than half the posterior standard deviation shown.

6. Decision rules

For determining the ELC in the context of species exchangeability, a number of decision rules, related to estimation of the HC_p for a specified value p of interest, have been proposed in the literature. We consider two existing rules and their generalisation to non-exchangeability under both M1 and M2. Generally, risk is measured/controlled via the 'potentially affected fraction' (PAF), the proportion of species whose tolerance lies below the ELC, with some intention to keep the PAF near or below $p\%$. The choice of p is seen to be a policy decision for the risk manager; the standard requirement is 5. However, the justification for this choice comes largely from some validation studies carried out afterwards to examine the consequences. A high PAF corresponds to a high risk for the assemblage of species.

6.1. Risk approaches for determination

We denote the proposed $\log_{10}(\mathrm{ELC})$ for a new substance by δ. In all the cases we consider, it can be shown (see Appendix A.2) that δ is of the form $\hat{\mu} - \kappa_p \hat{\sigma}$. Here, $\hat{\mu}$ and $\hat{\sigma}$ are natural estimates of μ and σ from the data for the new substance while κ_p does not depend on these data, although it does always depend on n and p and the risk measure. κ_p might be described as a standardised assessment shift so that $10^{\kappa_p \hat{\sigma}}$ is the variable assessment factor referred to in Section 2.2. Risk managers should find the rules appealing and transparent for reasons discussed later.

In all cases, $\hat{\mu}$ is the standard weighted least squares unbiased estimate of μ, obtained by correcting the measurement for the special species to remove the bias k and increasing its weight to allow for the reduction in variability implied by ϕ. Under M1, $\hat{\sigma}^2$ is simply the corresponding weighted least squares unbiased estimate of σ^2 whereas under M2 it is a weighted combination of that estimate and the prior mean for σ^2 implied by α and β. Consequently, on the original concentration scale the value determined for the ELC is a geometric mean of the adjusted toxicity data divided by the aforementioned variable assessment factor. The difference between M1 and M2 is that the latter stabilises the variability estimate $\hat{\sigma}$ by borrowing strength from the pool $\mathcal{G}_2$ of existing data; a corresponding adjustment is required to the value of κ_p which then depends on α.

Simple rules based on exchangeable versions of M1 were proposed by Aldenberg and Jaworska (2000) [AJ] and EFSA (2005) [EFSA]. The latter also considered the [EFSA] rule in the context of exchangeable M2; we determine the [AJ] version here for completeness (see Appendix A.2 for details). In what follows, note that $\mathrm{PAF}(\delta) = \Phi\big((\delta - \mu)/\sigma\big)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and we write $\mathrm{PAF}(\delta)$ to emphasise dependence on the decision rule.

The [AJ] approach is to demand high probability that $\mathrm{PAF}(\delta)$ is less than $p\%$. The risk manager specifies p, often taken to be 5 in practice, and a credibility requirement γ; the decision rule is to find δ so that γ is the probability that $\mathrm{PAF}(\delta)$ is less than $p/100$. Noting that $\mathrm{PAF}(\delta) \leq p/100$ if and only if $\delta \leq \log_{10}(\mathrm{HC}_p)$, δ satisfies

$$\mathrm{P}(\delta \leq \mu - K_p \sigma) = \gamma \tag{1}$$

where K_p is the $(100-p)$-th percentile of the standard normal distribution; the resulting κ_p depends on γ. The probability in (1) is computed with respect to the posterior distribution

of μ and σ for the new substance. It has been suggested by some that $\gamma = 0.95$ may be an appropriate choice (e.g. Wagner and Løkke 1991). However, current EU guidance (e.g. ECHA 2008a) requires results for $\gamma = 0.50$ to be presented along with those for $\gamma = 0.25$ and $\gamma = 0.75$.

The [EFSA] approach is to try to control PAF(δ) to be near some suitable value p% which the risk manager specifies. Then δ is the value for which the expected PAF is $p/100$ and so satisfies

$$\mathrm{E}\Big(\Phi\big((\delta - \mu)/\sigma\big)\Big) = p/100 \tag{2}$$

where again the expectation is with respect to the posterior distribution of μ and σ. The value of p will generally need to be smaller, for example $p = 1$, for the [EFSA] approach in order to achieve similar protection to that obtained by [AJ] with $p = 5$ when $\gamma = 0.95$.

To obtain the simple form $\delta = \hat{\mu} - \kappa_p \hat{\sigma}$, we have to assume that the hyper-parameters θ are known/specified precisely so that we actually compute the probability in (1) and the expectation in (2) using the posterior distribution of μ and σ conditional on θ. Consequences of uncertainty about θ are addressed in Section 7.

A number of features of these rules make them sensible and easy to apply: (i) each rule is easily computed and tables for κ_p can be produced for those who lack the necessary expertise or software (cf. Aldenberg and Jaworska 2000, Table 1, p. 5); (ii) each rule has the same form as in the exchangeable species case; (iii) the [AJ] rule is a Bayes rule under generalised absolute loss (Hickey et al., 2009); and (iv) the rules hold from the frequentist perspective in the sense that (1) and (2) remain valid if the calculations are with respect to the sampling distribution of the tolerance data for the substance, and also the sampling distribution of σ in the case of M2, instead of the posterior distribution of μ and σ.

6.2. Consequences of non-exchangeability

Application of revised decision rules will ultimately yield different consequences, but it is not immediately apparent to what degree. Figure 2 compares the values of δ obtained for each revised rule to those calculated under exchangeability for each substance in the $\mathcal{G}_1$ database; results are shown for $p = 5$ for each substance i in $\mathcal{G}_1$ for [AJ] ($\gamma = 0.50, 0.95$) and [EFSA]; we plot δ calculated under exchangeability versus the difference (to assist interpretation) between the values of δ obtained under non-exchangeability and exchangeability.

The horizontal dashed lines indicate where the decision rules are equal; points above the line indicate substances for which the revised ELC is higher than the original, i.e. where it is ecologically less conservative. An important observation for regulators is that the new rules, although correcting for a single sensitive species, do not necessarily lead to higher ELCs. In fact, the δ values based on non-exchangeability are higher than their exchangeable model versions for between 60% and 68% of assessed substances (Figure 2) for [AJ] ($\gamma = 0.50$) and [EFSA], and between 52% and 56% for [AJ] ($\gamma = 0.95$). This is due partly to the fact that although the offset hyper-parameter k is positive, the variance estimate also changes leading sometimes to higher *and* sometimes to lower values of δ. The largest differences occur for substances where the non-exchangeable decision rule is lower than the corresponding exchangeable version and under M1 this feature is more pronounced for [AJ] ($\gamma = 0.95$) as the change of model has more effect in the tails of the posterior distribution for the HC_5

There is some double counting of data here since the estimated hyper-parameters θ derive from the same database used to explore the consequences. However, the estimates are based on many substances and would change relatively little on omitting one. Moreover,

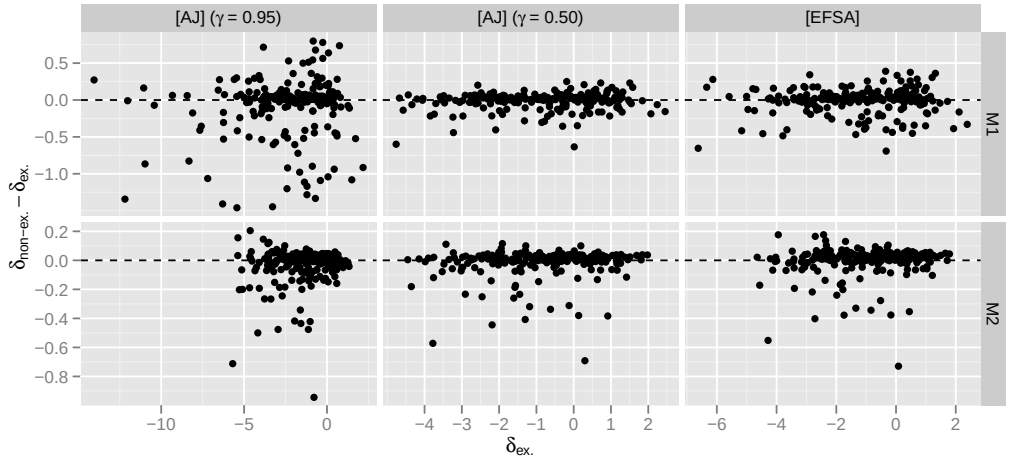

Fig. 2. Consequences of non-exchangeability for $p = 5$ for all substances in $\mathcal{G}_1$: δ derived under exchangeability versus the difference between δs derived under non-exchangeability and exchangeability

the estimates are those which will be used in the decision rules we propose for risk managers and it is the consequences of the change to those rules which we wish to evaluate.

7. Consequences of ignoring hyper-parameter uncertainty

In Section 6, we assumed that hyper-parameter uncertainty could safely be ignored, resulting in a simple form for the rules for determining the ELC. Here we seek to show that the rules derived still perform well even if we allow for hyper-parameter uncertainty. The simple form arose from solving (1) and (2) making the approximation of using the posterior distribution of μ and σ conditional on taking the hyper-parameters θ fixed at their MAP estimates in place of the marginal posterior distribution of μ and σ. Approximate numerical solution is possible when θ is uncertain but it is not easy to ensure reliability or accuracy.

However, the left-hand sides of (1) and (2) can each be seen as measuring performance of a chosen value of δ and the right-hand sides as specifying intended performance. For [AJ], the performance measure is the probability that the PAF is less than p; for [EFSA], it is the expected PAF. Consequences of ignoring hyper-parameter uncertainty for each decision rule may be assessed by taking δ fixed at the value used for each substance in producing the corresponding panel in Figure 2 and accurately computing the left-hand-side of (1) for [AJ] or (2) for [EFSA] in order to obtain *attained performance*. The result may be compared to the intended value: γ for [AJ] or p for [EFSA]. If an attained value is greater (or lower) than intended, ignoring hyper-parameter uncertainty has led to higher (or lower) than intended protection of the ecological community.

Computation of attained performance for each substance is simple once one has a large random sample of values from the posterior distribution of θ; one calculates the performance of δ for each value of θ and then averages. We took a Markov chain Monte Carlo sample of 10,000 values from the posterior density of the hyper-parameters under each behavioural

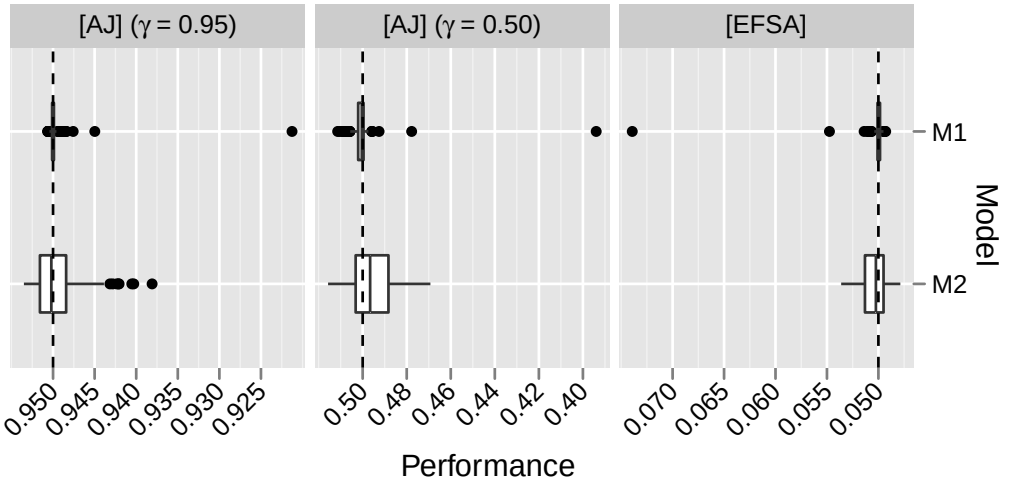

Fig. 3. Box-plots of per-substance attained performance for decision rules obtained ignoring hyper-parameter uncertainty. Attained performance is expected PAF for [EFSA] and credibility that PAF $<$ p% for [AJ] ($\gamma = 0.50$ and $\gamma = 0.95$), computed allowing for hyper-parameter uncertainty.

model, using a Metropolis random walk sampler with a normal proposal distribution based on the Laplace approximation to the posterior, which can be performed using regular statistical software; see for example Albert (2007, p. 110).

Figure 3 shows attained performance for each substance in $\mathcal{G}_1$ for both behavioural models with $p = 5$. The same three ELC rules are considered as in Figure 2: [AJ] ($\gamma = 0.95$), [AJ] ($\gamma = 0.50$) and [EFSA]. In each plot the intended performance level is emphasised by a dashed line. In interpreting differences between intended and attained performance, we must recognise that this is intermediate tier ERA, that the chosen value of $p = 5$ has no direct ecological meaning and that the actual PAF will always be highly variable between substances due to the relatively small numbers of species tested. With the exception of one substance, attained performance under M1 does not differ from intended performance in any practical sense; for example the difference between 50% credibility and 48% credibility is negligible. Even in the exceptional case, the difference may well be acceptable to risk managers. Under M2, there are somewhat larger typical differences between attained and intended performance but these are still tolerable in our opinion. In all cases, it appears that slight under-protection occurs more often than over-protection.

Earlier, we examined the sensitivity of hyper-parameter estimates to our choice of prior distribution for the hyper-parameters as we cannot be sure that our chosen prior is the best representation of prior knowledge. We also evaluated the attained performance for each substance of each δ shown in Figure 2 using the posterior distribution for μ and σ obtained using each of the alternative priors described in Section 5. Naturally, there were some differences between attained and intended performance. Nevertheless, for the majority of the alternative priors, the differences were small, especially under M1, and even in the worst case the differences were less than 20% of intended p for [EFSA] and of intended $1 - \gamma$ for [AJ]. In effect, the rules were still attaining the right magnitude of performance despite the

Table 4. MAP estimates under D1 for hyper-parameters k', ϕ', α' and β' with posterior standard deviations in parentheses.

	k'	ϕ'	α'	β'
M1	0.458 (0.060)	0.642 (0.076)	—	—
M2	0.452 (0.056)	0.604 (0.065)	1.52 (0.22)	0.315 (0.069)

fact that the original prior was being used for determining δ and the alternative priors for computing attained performance.

8. Comparison of models for non-exchangeability

In Section 4, we introduced our model for non-exchangeability and noted its tractability compared to the model proposed in EFSA (2005). We now consider the evidence in favour of one over the other from other perspectives. We denote by D1 the model introduced by EFSA (2005), with non-exchangeability hyper-parameters k' and ϕ' and by D2 our model with parameters k and ϕ. Details of D1 and D2 were provided in Section 4. There we did not distinguish ϕ from ϕ'; however, although apparently the same, ϕ' and ϕ have different meanings due to the difference between D1 and D2 in the treatment of the mean for the special species. Table 4 gives estimates under D1 corresponding to those under D2 given earlier in Table 3. In principle, under M2, estimates of α and β differ for D1 and D2 due to the different treatment of non-exchangeability; however the tabulated values coincide.

Suppose we take a substance out of the database $\mathcal{G}_1$ and consider it to be the substance under current assessment. We compare the two non-nested non-exchangeability models D1 and D2 for each substance using a Bayes factor (Bernardo and Smith, 1994; Kass and Raftery, 1995) to measure the evidence in favour of D1 against D2. The Bayes factor for a substance is the ratio of the marginalised likelihoods under D1 and D2 where each marginalised likelihood is the expectation, calculated using the prior distribution of μ and σ, of the conventional likelihood for the data for the substance. Evidence provided by a Bayes factor in favour of D1 or D2 may be interpreted using a descriptive categorisation such as that proposed by Kass and Raftery (1995, Section 3.2) which provides an intuitive and practical approach to model comparison for applied Bayesian statistics. Note that there are some technical issues when applying Bayes factors with improper prior distributions and we have to treat the hyper-parameters as fixed; details are given in Appendix A.3 along with the formula for the Bayes factor.

Figure 4 shows the Bayes factors for individual substances separately for M1 and M2. Under M2, all lie in a range deemed by Kass and Raftery (1995) not to indicate a significant advantage for either model. The same is true for most substances under M1 although there are a few in each direction strongly favouring D1 or D2. However, 131 of 220 Bayes factors are positive under M1 and 141 under M2 which may suggest some overall preference for D1.

A simple summary of the overall evidence for D1 against D2 is the overall Bayes factor, obtained as the product of the per-substance Bayes factors since substances are conditionally independent when θ is fixed. Under M1, this is 2.6 which Kass and Raftery (1995) describe as 'not worth a bare mention' whereas under M2 it is 426 which they consider 'decisive' in favour of D1. However, it is unclear how much ignoring hyper-parameter uncertainty undermines the calculation, especially given that the estimates are based on the same data. Unfortunately, there is little expert knowledge on which to base proper prior distributions and none which would prevent the Bayes factor from depending arbitrarily on the relative

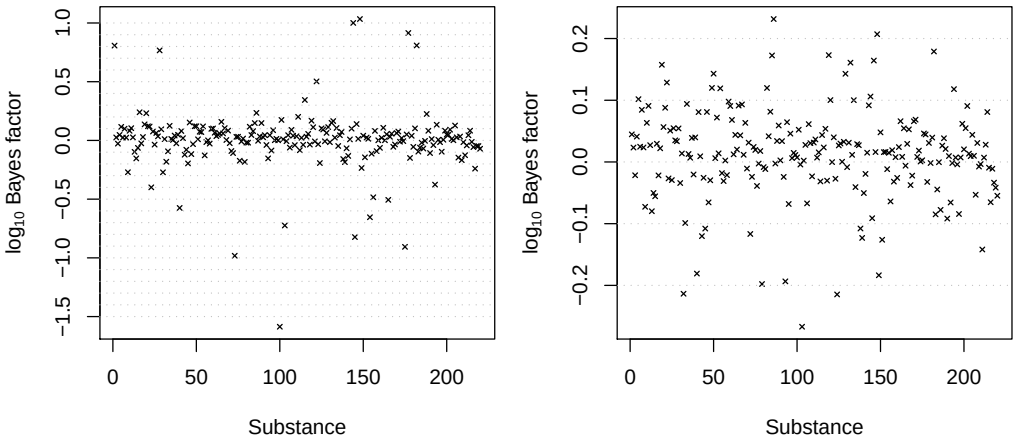

Fig. 4. Bayes factors for D1 versus D2 for substances in $\mathcal{G}_1$. Left: M1; Right: M2.

prior density of k and k'. We are left with the facts that: (i) D2 leads to tractable risk calculations, (ii) individual substances do not distinguish D1 from D2, (iii) the overall picture slightly favours D1 over D2 but only if the same form of non-exchangeability is assumed to hold throughout. D2 is our pragmatic choice.

9. Discussion

We have provided evidence to support a previous informal view that an important test species, *O. mykiss* (the rainbow trout), fails to satisfy the key exchangeability assumption in the SSD approach to ecotoxicology. We then showed how to adapt current modelling and procedures to allow for a single species with non-exchangeable tolerance, while retaining two key features: simplicity of decision rules and no need to share databases. However, the evidence clearly suggests that more than one species may be non-exchangeable.

In Section 2.3, we explained the difficulties in using the apparently natural approach of a crossed random effects model. In short, it would not lead to simple decision rules, it would require more sharing of data and would require careful reconsideration of the SSD concept, thereby violating our goal to seek procedures which would be sufficiently transparent to allow adoption by risk managers. We do not know if it would lead to better decision rule performance. Our solution has the merits that it addresses the problem of non-exchangeability for the standard test species, that it is a relatively straightforward adaptation of current methodology and that it seems to be reasonably well supported by data. Crucially, it is simple enough that risk managers need not radically alter their approach.

Mathematically, and to some extent computationally, it is straightforward to extend the model and decision rules in this paper to allow for multiple special species. However, this introduces two fundamental problems. The first is to decide which and how many species should be treated as having non-exchangeable tolerances. It is likely that disagreement on this issue would make it difficult to establish standard decision rules. The second, and more serious, conceptual problem is that the SSD is supposed to be a surrogate for ecosystems. In our current proposal, the SSD does not describe the special species and protection is still achieved purely in terms of the SSD although the special species contributes information.

In removing more species from the SSD, we would eventually have to consider how to use the SSD together with the special species' tolerances in order to achieve protection goals.

An alternative would be to model SSDs as mixtures (Grist et al., 2006; Hickey et al., 2008) where species in the ecological community are grouped taxonomically. While it wouldn't account fully for species non-exchangeability, it might be appropriate where sensitive groups are known to be measured. It has appeal for complex and diverse communities, but would need additional knowledge of taxonomic weightings, more data, and specialist statistical software for working with mixture distributions. Consequently, such models are unlikely to become commonplace tools for intermediate tier ERA.

Current ERA procedures generally use only the data for the substance under consideration. Decision rules based on hyper-parameters estimated from multi-substance databases may not immediately appeal to the user-community but at least do not require general sharing of databases. However, a conventional Bayesian approach would involve updating hyper-parameters as more data become available. That would require someone to augment databases and re-compute hyper-parameter estimates on an on-going basis. In our proposal, the hyper-parameters would be static and used over a significant period of time for many risk assessments. This is not intended to improve on the standard paradigm but is simply pragmatic. It removes the requirement for those actively involved in ERA to use sophisticated statistical software and allows users instead to use spreadsheet software and publishable look-up tables, since more complex analysis would only be performed occasionally by statisticians. There remains the issue of how and when databases would be updated but that is a problem for the ERA community and not for statisticians.

Acknowledgements

Hickey would like to thank the Engineering and Physical Sciences Research Council and The Food and Environment Research Agency for funding his research during his Ph.D.

References

Albert, J. (2009). *Bayesian Computation with R.* **1st ed.** New York: Springer.

Aldenberg, T., and Slob, W. (1993). Confidence Limits for Hazardous Concentrations Based on Logistically Distributed NOEC Toxicity Data. *Ecotoxicology and Environmental Safety*, **25**: 48–63.

Aldenberg, T. and Jaworska, J. S. (2000). Uncertainty of the Hazardous Concentration and Fraction Affected for Normal Species Sensitivity Distributions. *Ecotoxicology and Environmental Safety*, **46**: 1–18.

Aldenberg, T., Jaworska, J. S. and Traas, T. P. (2002). Normal Species Sensitivity Distributions and Probabilistic Ecological Risk Assessment. In: Posthuma, L., II Suter, G. W. and Traas, T. P. Eds. *Species Sensitivity Distributions in Ecotoxicology.* Boca Raton: Lewis Publishers, 49–102.

Aldenberg, T. and Luttik, R. (2002). Extrapolation Factors for Tiny Toxicity Data Sets From Species Sensitivity Distributions With Known Standard Deviation. In: Posthuma, L., II Suter, G. W. and Traas, T. P. Eds. *Species Sensitivity Distributions in Ecotoxicology.* Boca Raton: Lewis Publishers, 103–118.

637 Alexander, D. E. and Fairbridge, R. W. (1999). *Encyclopaedia of Environmental Science.*
638 Amsterdam: Kluwer Academic Publishers.

639 Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* Chichester: Wiley.

640 De Zwart, D. (2002). Observed Regularities in SSDs for Aquatic Species. In: Posthuma, L.,
641 II Suter, G. W. and Traas, T. P. Eds. *Species Sensitivity Distributions in Ecotoxicology.*
642 Boca Raton: Lewis Publishers, 133–154.

643 Duboudin, C., Ciffroy, P. and Magaud, H. (2004). Effects of Data Manipulation and Statis-
644 tical Methods On Species Sensitivity Distributions. *Environmental Toxicology and Chem-*
645 *istry*, **23**: 489–499.

646 Dwyer, F. J., Mayer, F. L., Sappington, L. C., Buckler, D. R., Bridges, C. M., Greer, I. E.,
647 Hardesty, D. K., Henke, C. E., Ingersoll, C. G., Kunz, J. L., Whites, D. W., Augspurger,
648 T., Mount, D. R., Hattala, K. and Neuderfer, G. N. (2005). Assessing Contaminant
649 Sensitivity of Endangered and Threatened Aquatic Species: I. Acute Toxicity of Five
650 Chemicals. *Archives of Environmental Contamination and Toxicology*, **48**: 143–154.

651 EC (European Commission) (2002). Guidance Document on Aquatic Ecotoxicology in the
652 Context of the Directive 91/414/EEC. SANCO/3268/2001, **4**: 62.

653 EC (European Commission) (2006). Regulation (EC) No. 1907/2006 of the European Par-
654 liament and of the Council of 18 December 2006. *O*fficial Journal of the European Union,
655 L 396/1.

656 ECHA (European Chemicals Agency) (2008a). Guidance for the Implementation of
657 REACH: *Guidance on Information Requirements and Chemical Safety Assessment Chap-*
658 *ter R.10: Characterisation of Dose [Concentration]-Response for Environment*. May 2008.

659 ECHA (European Chemicals Agency) (2008b). Guidance for the Implementation of
660 REACH: *Guidance on Information Requirements and Chemical Safety Assessment Chap-*
661 *ter R.19: Uncertainty Analysis*. May 2008.

662 EFSA (European Food Safety Authority) (2005). Opinion of the Scientific Panel on Plant
663 Health, Plant Protection Products and their Residues on a Request from EFSA Related
664 to the Assessment of the Acute and Chronic Risk to Aquatic Organisms with Regard to
665 the Possibility of Lowering the Uncertainty Factor if Additional Species were Tested. *The*
666 *EFSA Journal*, **301**: 1–45.

667 Forbes, V. E. and Calow, P. (2002a). Extrapolation in Ecological Risk Assessment: Bal-
668 ancing Pragmatism and Precaution in Chemical Controls Legislation. *BioScience*, **52**:
669 249–257.

670 Forbes, V. E. and Calow, P. (2002b). Species Sensitivity Distributions Revisited: A Critical
671 Appraisal. *Human and Ecological Risk Assessment*, **8**: 1625–1640.

672 Goldstein, H. (1995). *Multilevel Statistical Models.* **2nd ed.** London: Arnold.

673 Grist, E. P. M., O'Hagan, A, Crane, M., Sorokin, N., Sims, I. and Whitehouse, P. (2006).
674 Bayesian and Time-Independent Species Sensitivity Distributions for Risk Assessment of
675 Chemicals. *Environmental Science and Technology*, **40**: 395–401.

Hickey, G. L., Kefford, B. J., Dunlop, J. E. and Craig, P. S. (2008). Making Species Salinity Sensitivity Distributions Reflective of Naturally Occurring Communities: Using Rapid Testing and Bayesian Statistics. *Environmental Toxicology and Chemistry*, **27**: 2403-2411.

Hickey, G. L., Craig, P. S. and Hart, A. (2009). On The Application of Assessment Factors in Ecological Risk. *Ecotoxicology and Environmental Safety*, **72**: 293–300.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, **90**: 773–795.

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**: 508–515.

Posthuma, L., II Suter, G. W., Traas, T. P. Eds. (2002). *Species Sensitivity Distributions in Ecotoxicology.* Boca Raton: Lewis Publishers.

Raimondo, S., Vivian, D. N., Delos, C. and Barron, M. G. (2008). Protectiveness of Species Sensitivity Distribution Hazard Concentrations for Acute Toxicity Used in Endangered Species Risk Assessment. *Environmental Toxicology and Chemistry*, **27**: 2599–2607.

Rand, G. M. ed. (1995). *Fundamentals of Aquatic Toxicology: Effects, Environmental Fate and Risk Assessment.* **2nd ed.** Philadelphia: Taylor and Francis.

Stephan, C. E. (2002). Use of Species Sensitivity Distributions of Water Quality Criteria for Aquatic Life by the U.S. Environmental Protection Agency. In: Posthuma, L., II Suter, G. W. and Traas, T. P. Eds. *Species Sensitivity Distributions in Ecotoxicology.* Boca Raton: Lewis Publishers, 211–220.

Wagner, C. and Løkke, H. (1991). Estimation of Ecotoxicology Protection Levels from NOEC Toxicity Data. *Water Research*, **25**: 1237–1242.

A. Appendix

A.1. Parameter estimation

Here we give details of the posterior distributions for the hyper-parameters under M1 and M2. In the interest of clarity, we extend the notation of Section 4 by writing $\tau_i = 1/\sigma_i^2$ and we note that the transformed prior density is $p(\tau_i) \propto 1/\tau_i$ for $\tau_i > 0$. We also denote the database of toxicity data as $\mathbf{Y}$. The collection of $n_i - 1$ species tested with substance i, but not including the special species, is denoted J_i^*.

Under D2, for both M1 and M2, define

$$\hat{\mu}_i = \frac{\phi^{-2}(y_i^\dagger + k) + \sum_{j \in J^*} y_{ij}}{\phi^{-2} + n_i - 1}; \text{ and,} \tag{3}$$

$$\hat{\sigma}_i^2 = \frac{2\beta + (n_i - 1)\tilde{\sigma}_i^2}{2\alpha + (n_i - 1)}; \text{ where} \tag{4}$$

$$\tilde{\sigma}_i^2 = \frac{1}{n_i - 1}\left[\phi^{-2}(y_i^\dagger + k - \hat{\mu}_i)^2 + \sum_{j \in J^*}(y_{ij} - \hat{\mu}_i)^2\right], \tag{5}$$

where, for M1, $\alpha = \beta = 0$. Note the implicit dependence on hyper-parameters and also that $\hat{\mu}_i$ and $\tilde{\sigma}_i^2$ are the usual weighted least squares unbiased estimators of μ_i and σ_i^2. For M2, $\hat{\sigma}_i^2$ is also unbiased from the frequentist viewpoint if one incorporates drawing σ_i^2 from an inverse-gamma population of variances into the sampling scheme.

Under D2 and M1, writing $\mu_{\mathcal{G}_1}$ and $\tau_{\mathcal{G}_1}$ as shorthand for the vectors of the μ_i and τ_i for $i \in \mathcal{G}_1$ respectively, and v_t for the number of substances in $\mathcal{G}_t$ $(t = 1, 2)$, we easily obtain the likelihood function for all the unknown parameters:

$$\begin{aligned} L(k, \phi, \mu_{\mathcal{G}_1}, \tau_{\mathcal{G}_1}) &\propto \prod_{i \in \mathcal{G}_1} \phi^{-1} \tau_i^{n_i/2} \exp\left\{ -\tfrac{1}{2}\tau_i \left[\phi^{-2}(y_i^\dagger - \mu_i + k)^2 + \sum_{j \in J_i^*} (y_{ij} - \mu_i)^2 \right] \right\} \\ &= \phi^{-v_1} \prod_{i \in \mathcal{G}_1} \tau_i^{n_i/2} \exp\left\{ -\tfrac{1}{2}\tau_i \left[(\phi^{-2} + n_i - 1)(\hat{\mu}_i - \mu_i)^2 + (n_i - 1)\hat{\sigma}_i^2 \right] \right\} \end{aligned}$$

Multiplying by the joint prior density defined in Section 5 for k, ϕ, μ_i and τ_i $(i \in \mathcal{G}_1)$ yields the un-normalised posterior distribution, and after integration with respect to each μ_i and τ_i, we obtain the posterior density for k and ϕ:

$$p(k, \phi \,|\, \mathbf{Y}) \propto \phi^{-v_1} \prod_{i \in \mathcal{G}_1} \frac{\Gamma(\hat{\alpha}_i)}{\hat{\beta}_i^{\hat{\alpha}_i}} \frac{1}{\sqrt{\phi^{-2} + n_i - 1}}, \tag{6}$$

where $\hat{\alpha}_i = \frac{1}{2}(n_i - 1)$ and $\hat{\beta}_i = \hat{\alpha}_i \hat{\sigma}_i^2$. Maximising this function with respect to its arguments subject to the constraint $\alpha = \beta = 0$ determines the joint MAP estimator for k and ϕ.

Under D2 and M2, we use the additional $v_2 - v_1$ substances in $\mathcal{G}_2 \backslash \mathcal{G}_1$ and estimate α, β, k and ϕ. Momentarily continuing to treat the τ_i as parameters, the likelihood is now

$$L(k, \phi, \mu_{\mathcal{G}_1}, \tau_{\mathcal{G}_1}) \prod_{i \in \mathcal{G}_2 \backslash \mathcal{G}_1} \tau_i^{n_i/2} \exp\left\{ -\tfrac{1}{2}\tau_i \left[n_i(\bar{y}_i - \mu_i)^2 + (n_i - 1)s_i^2 \right] \right\}$$

where $\mu_{\mathcal{G}_2 \backslash \mathcal{G}_1}$ and $\tau_{\mathcal{G}_2 \backslash \mathcal{G}_1}$ are similarly defined as per earlier, and $\bar{y}_i$ and s_i are the sample mean and standard deviation of y_{ij} $\forall j \in J_i$. Now, we must multiply by the sampling density, $p(\tau_i \,|\, \alpha, \beta) = [\beta^\alpha / \Gamma(\alpha)] \tau_i^{\alpha - 1} e^{-\beta \tau_i}$ for $i \in \mathcal{G}_2$, recalling $\mathcal{G}_1 \subseteq \mathcal{G}_2$ and integrate with respect to each $\tau_i > 0$ to obtain the true likelihood under M2. However, we then intend to multiply by the prior density $p(k, \phi, \alpha, \beta, \mu_{\mathcal{G}_2}) \propto 1$ and integrate with respect to each μ_i to obtain the marginal posterior and it is easier to reverse the order of integration (as earlier) to obtain

$$p(\alpha, \beta, k, \phi \,|\, \mathbf{Y}) \propto \left[\frac{\beta^\alpha}{\Gamma(\alpha)} \right]^{v_2} \phi^{-v_1} \left(\prod_{i \in \mathcal{G}_2} \frac{\Gamma(\tilde{\alpha}_i)}{\tilde{\beta}_i^{\tilde{\alpha}_i}} \right) \left(\prod_{i \in \mathcal{G}_1} \frac{1}{\sqrt{\phi^{-2} + n_i - 1}} \right), \tag{7}$$

where $\tilde{\alpha}_i = \alpha + \hat{\alpha}_i$ and $\tilde{\beta}_i = \beta + \hat{\beta}_i$ for $i \in \mathcal{G}_1 (\supseteq \mathcal{G}_2)$.

Under D1, $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ in (3) and (4) are now functions of τ_i as k must be replaced by $k'/\sqrt{\tau_i}$ and we also replace α by α', β by β' and ϕ by ϕ'. Consequently, when calculating the equivalent of (6) and (7), the integrals with respect to μ_i can still be done in closed form but integration with respect to τ_i must be approximated numerically .

A.2. Decision rules under D2

For M2, it is a straightforward generalisation of standard Bayesian calculations for normal sampling to obtain the posterior distribution of μ and σ^2 — the parameters of an SSD for

a new substance — conditional on known θ and tolerance measurements for a substance: $1/\sigma^2$ has a gamma distribution with shape $\tilde{\alpha} = \alpha + \frac{1}{2}(n-1)$ and mean $1/\hat{\sigma}^2$ and, given σ, μ has a normal distribution with mean $\hat{\mu}$ and variance $\sigma^2/(\phi^{-2}+n-1)$, given by (3) and (4) respectively after dropping the subscript i. Under M1, $\hat{\sigma}^2$ simplifies to $\tilde{\sigma}^2$.

Decision rules are determined to be of the form $\hat{\mu} - \kappa_p\hat{\sigma}$ for both [AJ] and [EFSA] methods. This follows from two standard results for the normal-inverse-gamma posterior distribution for μ and σ^2: (i) $\mu - K_p\sigma$ has a re-scaled non-central t-distribution; and (ii) the predictive distribution of a further observation is a re-located and re-scaled t-distribution. For [AJ], the decision rule follows directly from (i), while for [EFSA], one needs to note that E(PAF(δ)) is the probability that the tolerance of a random species lies below δ, which is given by (ii).

For the [AJ] rule, $\psi\kappa_p$ is the γ-th percentile of the non-central t-distribution with $\eta = 2\alpha + n - 1$ degrees of freedom and non-centrality parameter ψK_p, where $\psi^2 = \phi^{-2} + n - 1$ is the total weight of the observations. For [EFSA], $\kappa_p/\sqrt{1+\psi^{-2}}$ is the $(100-p)$-th percentile of the (central) t-distribution with η degrees of freedom. Note that κ_p values differ for M1 and M2 and are non-comparable as they are to be applied to different estimates of σ. For M1, take $\alpha = \beta = 0$. Similarly, calculations under exchangeability may be recovered by taking $k = 0$ and $\phi = 1$.

A.3. Bayes factors

For Bayes factors for D1 against D2 for a new substance, first consider M2. Let (k', ϕ') and (k, ϕ) denote the estimated values of the non-exchangeability hyper-parameters under D1 and D2 respectively and let (α', β') and (α, β) be the respective variance heterogeneity parameters. We take the hyper-parameters to be fixed in each mode because Bayes factors are generally undefined when improper priors are used and also because, as in Section 6.2, the models we propose for actual use have fixed hyper-parameters. Next, recall that our prior distribution for μ is the improper uniform distribution on the real line so that we may exploit (7) to obtain the terms for a single substance under D2. With the form of the likelihood function given in Appendix A.1, we obtain the terms for a single substance under D1, upon which we can see that the Bayes factor in favour of D1 over D2 is

$$\frac{{\beta'}^{\alpha'}}{\beta^{\alpha}}\frac{\Gamma(\alpha)}{\Gamma(\alpha')}\frac{\phi\sqrt{\phi^{-2}+n-1}}{\phi'\sqrt{{\phi'}^{-2}+n-1}}\frac{\tilde{\beta}^{\tilde{\alpha}}}{\Gamma(\tilde{\alpha})}\int_0^{\infty}\tau^{\tilde{\alpha}'-1}\exp\{-\tfrac{1}{2}\tau[2\beta' + (n-1)\hat{\sigma}^2(\tau)]\}\,d\tau, \qquad (8)$$

where $\tilde{\alpha}$ and $\tilde{\beta}$ are defined as underneath (7) in Appendix A.1 (omitting the subscript i), $\tilde{\alpha}' = \alpha' + \hat{\alpha}$, and $\hat{\sigma}^2(\tau)$ is given by (4) and (5) (omitting the subscript i) with k replaced by $k'/\sqrt{\tau}$, α by α', β by β' and ϕ by ϕ'. The integral may be evaluated straightforwardly by numerical quadrature to high accuracy. The Bayes factor for M1 is given by (8), omitting the term ${\beta'}^{\alpha'}\Gamma(\alpha)/\beta^{\alpha}\Gamma(\alpha')$ and taking $\alpha' = \alpha = 0$ and $\beta' = \beta = 0$ in the remainder.

The prior distributions on μ and σ for M1 and μ for M2 are improper. However, following Bernardo and Smith (1994, p. 422), we argue that the Bayes factors are well defined as these parameters are identically operationally defined under D1 and D2 with respect to a hypothetical infinite population of exchangeable species in the SSD. In such contexts the Bayes factor obtained may be viewed as a limit of the one obtained using the same proper prior in the numerator and denominator