APPENDIX E (Updated March 2013)
Model Averaging Analysis of PESC and Elphick et al. (2011) data conducted by Dr. Carl Schwarz

An assessment of the effect of hardness on the dose-response curves to sulphates through the use of model averaging.

Carl James Schwarz (P.Stat. (Can), PStat (USA))
Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, BC V5A 1S6

Summary

Model averaging is method to determine the relative support for various hypotheses (models) about the effect of hardness on the dose-response curve to changing levels of sulphate. For example, one hypothesis may be that the dose-response curve against sulphate concentration is invariant over different hardness levels; or that the dose-response against sulphate is different over the hardness levels. The Akaike Information Criteria (AIC) measures the tradeoff between model fit (how well do the dose-response curves match the observed data) and model complexity (how many parameters are needed to describe the curve). The AIC provides a way to determine the relative weight to be attached to various models given by the observed data. Furthermore, the model weights can be used to provide estimates of benchmark doses (BMD, such as LCxx) that weight the estimates provided by the different models according to their support from the observed data.

The model averaging paradigm was applied to mortality and growth responses over a number of species collected in dose-response experiments at various levels of hardness. For the majority of responses, the majority of model weight was given to models where the dose-response curve is different at the hardness levels tested in the experiments. This implies that each hardness level will require a separate experiment to estimate the dose-response curve as there is no "sharing of information" across the hardness levels.

In the case of Fathead Minnows, the support was split between a model where the dose-response curves were different across hardness levels and a model where the dose-response curves have the same general shape, is parallel at the different hardness levels, and the dose-response curve shifted monotonically as hardness increased. The monotonic-shift model implies that once the dose-response curve is established for any hardness level, the same general curve can be used for higher hardness levels with a simply shift to the right as hardness increases.

In two cases (Lemna frond growth and final weight) there was substantial support for models where the dose-response curve was invariant across hardness levels. In this case,

the data are insufficient to distinguish between dose-response curves at different hardness levels. Natural variation in the data and a limited response to sulphate made it difficult to determine the effect of hardness on the dose-response curve.

Finally, in one case (Rainbow Trout mortality) support was almost event split between a model where the dose-response curve was invariant to hardness and where the dose-response curve was parallel and monotonically shifted to the right as hardness increased. In this case, extra-binomial variation was detected which made it difficult to distinguish the effect of hardness on the dose-response curve.

The general conclusion is that there is strong evidence that the dose-response curve varies by hardness for the majority of cases tested in this report. Only in those experiments where there was only a very limited response to sulphate (Lemna frond growth and final weight) or where there was evidence of excess natural variation (Rainbow Trout mortality) was there any substantial support for models where hardness appeared to have no effect on the dose-response curve.

For all experiments, model averaged estimates of the LC/IC10, LC/IC25, and LC/IC50 were obtained. These estimates are formed as a weighted average of the estimates from the individual models and the model averaged standard error accounts for both the uncertainty in the estimates for each individual model and the variation in estimates among the models examined.

1. Introduction

Dose-response studies are often used to estimate the risks associated with exposure to environmental hazards. For example, the B.C. Ministry of Environment recently commissioned a large study to investigate the response (mortality, growth, biomass) to sulphates under different water hardness from a number of species (e.g. Rainbow Trout, Daphnia).

In each study (combination of species, response, and hardness level), the dose-response curve is estimated using a statistical model that relates the dose of sulphate to the response. For example, a probit regression model may be used when the response is mortality; a log-logistic model may be used if the response is biomass. The fitted curve is then used to estimate a benchmark dose (BMD) such as the LCxx (the dose at which xx% additional mortality occurs over baseline mortality) or the ICxx (the dose at which the response (e.g. weight) is reduced from baseline). The estimates of BMD are model based because a direct estimation of the BMD may require using several hundred or thousands of organisms at a wide range of doses.

Many software packages (e.g. CETIS) provide a large number of dose-response curves (models) that may be fit to the same set of data, and each curve leads to a different estimate of the BMD. The risk manager is then faced with the problem of deciding which

model should be used, or equivalently, how to incorporate uncertainty in the BMD from different models that may fit the data equally well.

The analyst could choose the model that leads to the lowest BMD under the belief that this provides a conservative estimate of the BMD. Or, the analyst could choose the single best fitting model and use the associated BMD. Both of these strategies have the flaw that slight changes in the data or a new set of models could lead to a different "best" model being selected. As well, the selection of the "best fitting" model depends on the criteria used to define the fit of the model to the data. For example, the criteria used to select the best fitting model could be the model with the largest likelihood value; the model with the smallest total discrepancy between the data points and the fitted model (i.e. minimum sum-of-squares); the model with the smallest largest-discrepancy (i.e. the mini-max criteria), etc. D and different criteria could lead to different choices of the best model.

This report discusses a third option, model averaging, where the BMDs are "averaged" based on the support each model provided by the data. Burnham and Anderson (2002) and Anderson (2008) provide a comprehensive reference on this approach. Bailer et al (2005), Bailer, Noble and Wheeler (2005) provide examples applied to risk assessment using dose-response models.

2. Studies used.

There are two groups of studies used in this report, generally called the Enviroment Canada (EC) studies and the Nautilus Environmental (NA) studies (named after the organization that performed the studies).

In the Environment Canada studies, typically three water hardness values were tested on various freshwater species of aquatic organisms. The tests were done at a low water hardness (50 mg/L), a medium water hardness (100 mg/L), and a high water hardness (250 mg/L). Details of the experimental prot0col are found in Buday and Schroeder (2011). We use the data from Buday and Schroeder (2011) to assess if there is evidence of an effect of water hardness on the dose-response relationship between sulphate and the various endpoints measured.

Raw data were provided as an Excel workbook, the raw output sheets (in pdf format) from the analyses done by Buday and Schroeder (2011) using the CETIS software, and additional pdf files from the Saskatchewan Research Council who performed some of the work under sub-contract from Environment Canada.

In the Nautilus Environmental (NA) study, there were between one and four hardness levels (ranging from 15 to 320 mg/L) and a variety of freshwater species of aquatic organisms. We use the data from Elphick et al. (2010) to also assess if there is evidence of an effect of water hardness on the dose-response relationship between sulphate and the various endpoints measured. Only those organisms where at least two levels of water hardness were studies are used in this paper.

The raw data was extracted from copies of the raw output sheets (in pdf format) from the analyzes done by Elphick et al. (2010) using the CETIS software. Only the organisms where at least two water hardness levels were tested were used.

It is assumed that all the data presented are valid and no examination of the raw data for outliers or other anomalous points has been done.

The sampling protocol for each aquatic organism is presented in detail in Buday and Schoreder (2011) and Elphick (2010). A brief summary is presented in Table 1. All tests were performed at various levels of hardness of water and usually five or six nominal concentrations of sulphate. In the Environment Canada studies, the actual sulphate concentration was measured at the start and the end of the experiment and the average of the two values was used as the actual sulphate concentration. In the Nautilus studies, the nominal sulphate levels as recorded on the CETIS sheets were used directly. In Elphick et al. (2010, Table 2), a comparison of the measured vs. nominal sulphate levels showed a relatively good agreement. Most experiments also had a control (nominal zero concentration) of sulphate.

3. Theory of Model averaging

"All models are wrong, but some are useful" (Box and Draper, 1987) is an apt description for statistical modeling of many biological systems. While a Probit (see below) model may be an adequate approximation to the underlying dose-response relationship, the Probit model is not the "truth" and is necessarily "wrong" as no biological system follows such a simple dose-response curve. Consequently, estimates of BMDs are model dependent and reported precision (e.g. se) are conditional on the choice of this (wrong) model!

There are often several models that give essentially the same fit to the data, but could give rise to different estimates of the BMD. Model averaging (e.g. Burnham and Anderson, 2002; Anderson, 2008) is a way to recognize that all models are only approximations to reality and that there may be different models giving different answers. The basis behind model averaging is the use of AIC (Akaike Information Criteria)[1] which measures a combination of model fit and complexity. For example, if two models give the same fit to the data, but one model requires 2 parameters and the other model requires 50 parameters, then "Ockham's Razor" says that the model with fewer parameters is preferred. Similarly, as you increase the number of parameters, the fit of a model must improve (more parameters give a more flexible model), but is the improvement

[1] There are other model averaging criteria such as BIC (Bayesian Information Criteria). The same general principles apply to these other criteria.

"worthwhile" in light of the increase in complexity of the model. AIC[2] values are a function of likelihood (model fit) and number of parameters (complexity)

$$AIC = -2 \times \log(likelihood) + 2 \times (\#\ parameters)$$

Models with (arithmetically) smaller AIC values are preferred[3]. The actual numerical value of AIC is not important (nor interpretable), but the differences in AIC among competing models are important and lead to a measure of relative support for these models. Once a set of models is fit to the data, the AIC is computed for each model, and the models are sorted by the AIC values from the model with the best support from the data (lowest AIC) to the model with the worst support from the data (highest AIC). The difference in AIC (termed delta AIC) is found as:

$$\Delta AIC_{\text{model}} = AIC_{\text{model}} - AIC_{\text{best model}}$$

[By definition, the $\Delta AIC = 0$ for the model with the best support.] If two models have similar AIC values (usually within 2 or 3 units of each other), then these models are similar in their support by the data. Models that differ by more than 5 AIC units from the model best supported by the data are usually not thought of as being competitive. AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. This penalty discourages overfitting (increasing the number of free parameters in the model improves the goodness of the fit, regardless of the number of free parameters in the data-generating process).

Model weights are computed for each model based on a normalized function of the ΔAIC

$$w_{\text{model}} = \frac{\exp(-\Delta AIC_{\text{model}} / 2)}{\sum \exp(-\Delta AIC_{\text{model}} / 2)}$$

These model weights range from 0 to 1 and sum to 1 over the models considered.

These model weights provide a way to combine BMDs over competing models. Each model provides an estimate of the BMD and a weighted average (based on the model weight) is the "best" estimate for this BMD:

$$\overline{BMD} = \sum_{\text{models}} w_{\text{model}} BMD_{\text{model}}$$

The SE from each model can also be combined and if the estimates of the BMD vary considerably among models, an extra component of variation to account for this variation in estimates of BMD is also included:

$$se(\overline{BMD}) = \sum_{\text{models}} w_{\text{model}} \sqrt{\left(se_{\text{model}}\right)^2 + (BMD_{\text{model}} - \overline{BMD})^2}$$

---

[2] AIC values are usually corrected for small sample sizes. This is not detailed in this report but has been done in the examples that follow. The small sample corrected version of AIC is referred to as AICc.

[3] For example a model with an AIC of -10 is preferred over a model with an AIC of -5.

Confidence intervals are formed in the usual way based on the estimate and its standard error. For example, an approximate 95% confidence interval is found as

$$\overline{BMD} \pm 1.96 se(\overline{BMD})$$

The AIC paradigm is quite different from hypothesis testing and p-value approaches. The hypothesis testing and p-value approaches assume that one of the two models under consideration is correct. This is not biologically supported. For example, suppose you are interested in the effect of water hardness on the dose-response curve. The null hypothesis says that there is NO effect of water hardness. This hypothesis is clearly biologically wrong -- there must be an effect of water hardness - it may be small and not detectable from the data, but the hypothesis of no effect is "wrong" on biological grounds as there are always effects (do you really think that the mortality rates at different hardness levels are identical to 20 decimal places?). The p-value approach uses this (unrealistic) hypothesis of no-effect merely as a “straw-man” against which the data are examined to see if the data provide evidence against this “straw-man.” The AIC paradigm recognizes that all models are wrong, and provides a way to "quantify" if the fit is adequate compared to a more complex model.

Models do not have to be "nested" or of the same type to use the AIC paradigm. For example, the model set for a mortality dose-response curve could include a Probit model, a logistic model, a Gompertz model etc. However, all models must be likelihood based, so non-parametric models (such as non-parametric Karber-Spearman method for finding the LC50) are not directly usable. Likelihood methods are a scientifically defensible way to fit statistical models that uses all of the information in the data. Many existing methods are likelihood models in disguise (e.g. least squares for linear regression with normally distributed data is a likelihood fit).

One key assumption of the AIC paradigm is that the models chosen in the model set are sensible approximations to reality. For example, if the data showed an increasing effect of dose on mortality and only models that allowed for a decreasing effect of dose on mortality are fit, AIC will still rank these silly models and give the relative ranking of these silly models. Consequently, the fit of the model should also be ascertained by the analyst (this usually is done via residual plots and other methods).

The models in the model set should be specified in advance and the temptation to “data dredge” should be avoided. “Data dredging” would involve looking at the data and adding models that fit this particular dataset well, but have no a priori biological rationale. The danger is that the added models based on inspection of the data may be a good fit for this particular set of data, but minor changes in the data set would lead you to choose a different model to add. In reality, some model specification is driven by a preliminary look at the data, e.g. is hormesis present, and as along as the general class of models added is very general, this should be acceptable.

Because BMDs are typically computed as a function of the model parameters, there is some ambiguity in how the BMDs should be averaged. For example, should the BMD be averaged on the log-scale and then the averages are anti-logged, or should the averaging

take place directly on the anti-log scale. There is no biological definitive answer (for example, concentration of sulphates are measured on mg/L scale, but pH are measured on a logarithmic scale). The two approaches can lead to slightly different answers because the log() function is a non-linear transform, but the two methods should lead to similar results. As many models used in this project fit models where sulphates are measured on the log() scale, the model averaging will take place on the log-scale with a final anti-log() taken at the end of the process. This will lead to asymmetric confidence intervals on the anti-log scale. For example, from Table 4, the estimated BMD on the log-scale is 4.84 (SE 0.51). This gives 95% confidence intervals on the log-scale of (3.85, 5.84) which lead to 95% confidence intervals on the antilog scale of $\left(e^{3.85}=47, e^{5.84}=342\right)$.

Wheeler and Bailer (2007) discuss an alternate way to use model averaging where the dose-response curves are model averaged and the model-averaged curve is used to find the BMD, rather than model averaging the BMDs directly. This approach has not been applied in this report.

In some cases, a model may fit the observed data reasonably well, but is unable to provide an estimate of the BMD. This usually happens for one of two reasons. First, the model should (in theory) provide an estimate of the BMD, but sparsity in the data leads to a model fit where the BMD not longer exists. For example, the mortality in the observed dose range in the study is relatively constant but because of natural variability, the observed mortality declines with dose (e.g. 2/10 die at dose 100, 1/10 die at dose 200, and 0/2 die at dose 300). A fitted Probit model would lead to a model where the mortality declines as function of dose and would never reach 50% mortality (the LC50) and so no estimate of the BMD is available.

Second, the model may fit the observed data well, but cannot be extrapolated outside the observed range of the data. This is most common with isotonic models where the isotonicity constraint can be applied within the observed range of the data, but it is unclear how to extrapolate outside the observed range of doses. For example, suppose that in the observed range of doses, the mortality rate ranged from 0% to 40% (at the highest observed dose). It is not clear how to estimate the LC50 as this endpoint is outside the range of the observed doses. Al that is known is that estimate of the LC50 is higher than the observed dose, but no estimate is available.

It is valid to include models where no BMD can be determined in the model set and to obtain a model weight for this model. This is a valid comparison of competing models in a general sense – which models are supported by the data. For specific BMDs, the model may or may not be able to provide an estimate (e.g. it may be able to provide an estimate of the LC10, but not of the LC50). In cases where the model cannot provide an estimate of a BMD, it is assigned a model weight of 0 (even though its weighting in the model selection may be higher). This is not contradictory as the two analyses are answering two different questions (1) which model is the best tradeoff in fit and complexity for the given data and (2) how much credence should be given to each models estimate of the BMD. Of course, cases where all of the high ranking models fail to provide estimates of the BMD while the low ranking models are able to provide estimates of the BMD indicate

more serious problems with the study – most likely the BMD is well outside the range of the observed data and extrapolations may be pure fiction!

4. Models used.

There are two classes of responses in this study – quantal responses where the mortality of organisms is measured as a function of dose, and continuous responses (e.g. biomass) measured as function of dose.

4.1 Mortality Responses.

For the mortality responses, both Probit (Bliss, 1934) and Logit (Berkson, 1944) models were used.

The basic Probit model assumes that the number of deaths follows a binomial distribution where the probability of mortality is "linked" to a linear function through the normal distribution. For example, consider the Probit model for a fixed hardness level – the statistical model is:

$$Dead_{ij} \sim Binomial(BatchSize_{ij}, p_i)$$

$$p_i = \Phi\left(\beta_0 + \beta_i \log\left(D_{ij}\right)\right)$$

where $Dead_{ij}$ is the number of dead organisms observed in the j^{th} batch out of the initial $BatchSize_{ij}$ units on tests at dose level (sulphate) D_i; β_0, β_1 are the intercept and slope in the Probit model; and Φ is the cumulative normal distribution. [The original papers on Probit analysis added 5 to the linear functions to avoid negative numbers in hand computations, but this is no longer required when using computers.] The parameters are estimated using maximum likelihood (e.g. via Proc Probit in SAS). Estimates of the LCxx values (i.e. at what concentration will a fraction xx or organism die) can be found once estimates of the slope and intercept are found by solving the equation

$$LCxx/100 = \Phi\left(\hat{\beta}_0 + \hat{\beta}_i \log\left(D\right)\right)$$

Maximum likelihood estimates are asymptotically the best possible estimates and extract the maximum amount of information from the data. Estimates of precision (i.e. standard errors) can be found automatically for the parameters of the likelihood equations and by the delta method (Taylor series expansion) for the LCxx values.

The formulation above assumes that the probability of death will decline to zero as the sulphate dose declines to 0. Probit models have been developed to deal with non-zero natural responses. In the original papers, the observed mortality at control doses was treated as a fixed known natural response and the Probit analysis applied only to mortalities above this level. This approach ignored the uncertainty in the estimate and the resulting estimates and standard errors from the remainder of the fit did not account for this. A more modern approach is to let the natural response rate be another parameter to

be estimated in the model along with the slope and intercept of the Probit function. Again consider the Probit model for a fixed hardness level – the statistical model is:

$$Dead_{ij} \sim Binomial(BatchSize_{ij}, p_i)$$

$$p_i = NR + (1 - NR)\Phi\left(\beta_0 + \beta_i \log\left(D_{ij}\right)\right)$$

where NR is the natural response (mortality) at no (the control batches) sulphate, i.e. the fraction of units expected to die in the absence of an effect of sulphate. The parameters are again estimated using maximum likelihood (e.g. Proc Probit in SAS).

Note that in models with a very small dose-response effect, there is some ambiguity in the parameterization. This is because it is very hard then to distinguish between a natural response, or a model with a slope close to 0 as both will give similar fits to the data. In cases like this, it may be better drop the natural response terms.

Because of the natural response, estimation of the LCxx values must be done with care. For example, the LC25 values refer to the dose that results in a 25% mortality of the organism that *survive the natural response*. Suppose that the estimated natural response is 13%. Consequently, only 87% of the organisms would survive in the absence of sulphates. The LC25 refers to the additional 25% of 87%=22% mortality above the natural response for a total mortality of 12% + 22% = 35%. The estimated LC25 value is found by now solving:

$$.35 = .13 + .87\Phi\left(\hat{\beta}_0 + \hat{\beta}_1 \log(D)\right)$$

which again leads to

$$.25 = \Phi\left(\hat{\beta}_0 + \hat{\beta}_1 \log(D)\right)$$

i.e. the LC25 does not correspond to the by dose which leads to an overall .25 mortality.

In some cases, the LCxx values cannot be estimated. For example, if the probit model has an estimated slope < 0, then the predicted mortality rate declines with dose. [A non-positive estimate of the slope typically occurs with sparse data where, just by chance, fewer mortalities occurred at higher doses than at lower doses.] Even if the Probit model does fit, the dose-response curve may be so shallow that the estimated LCxx value is well beyond the range of the observed doses in the study. For example, suppose that mortality rates range from 0 to 10% in the range of doses in the study. The estimated LC50 value will be far to the right of the observed doses. Extrapolation well beyond the observed range of doses may be inadvisable – consequently, any LCxx value that is more than 2x the maximum observed dose in the study is "deleted".

A goodness-of-fit statistic of the Probit model (both with and without a natural response) to the data is found by comparing the observed and expected counts:

$$X^2 = \sum \frac{\left(Dead_{ij} - BatchSize_{ij}\,\hat{p}_{ij}\right)^2}{BatchSize_{ij}\,\hat{p}_{ij}} + \sum \frac{\left(Alive_{ij} - BatchSize_{ij}\left(1 - \hat{p}_{ij}\right)\right)^2}{BatchSize_{ij}\left(1 - \hat{p}_{ij}\right)}$$

where $\hat{p}_{ij}$ is the predicted probability of death for each batch. If the assumptions of the model are satisfied, this statistics should follow a χ^2_{df} distribution where the *df* is found

appropriately. If the X^2 statistic is extreme, it indicates a lack-of-fit. There are two common reasons for lack-of-fit. First, the model itself can be wrong (e.g. the response is not linear on the Probit scale), or the structural model is valid (i.e. the response is linear on the Probit scale), but the data are more variable than expected from a binomial response. The latter is termed overdispersion. For example, consider the sample proportion of organisms that die in batches of 30 organisms where the underlying mortality rate is 30%. Statistical theory indicates that under the binomial model, the average number that would die would be 9 = 30(.3), but the actual number that could die would range from 4 to 14. If the observed number that dies ranged from 1 to 17, this would indicate overdispersion, even though the average number that dies is still be 9. Typically causes of overdispersion are non-independence in the fate of the organism. For example, if all the organisms are placed in the same test tube, a local contaminant could reduce/increased the survival rate of this batch from the projected 30%.

The consequence of overdispersion is that estimates remain unbiased, but the reported standard errors (and p-values derived from them) are understated, i.e. the results appear to be more precise than they really are.

Corrections for overdispersion were incorporated directly in the model through the random effect Probit models (Gibbons et al, 1994; Gibbons and Hedeker, 1994). In the random effect model, latent (unobserved) random noise is added to the Probit function:

$$Dead_{ij} \sim Binomial(BatchSize_{ij}, p_i)$$
$$p_i = NR + (1 - NR)\Phi\left(\beta_0 + \beta_i \log\left(D_{ij}\right) + \varepsilon_{ij}\right)$$
$$\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$$

for non-control doses of sulphate, and

$$Dead_{ij} \sim Binomial(BatchSize_{ij}, p_i)$$
$$p_i = \Phi\left(\Phi^{-1}\left(NR\right) + \varepsilon_{ij}\right)$$
$$\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$$

for control doses of sulphate, where ε_{ij} is a latent random effect that comes from a normal distribution with mean 0 and variance σ^2, i.e. adding extra variation in the mortality rate at a specified dose. So even if the expected mortality rate at a particular dose is 30%, the random effect (applied at the batch level) could vary this higher or lower. This model can also be fit using maximum likelihood (e.g. Proc Nlmixed in SAS). Estimates from the fitted model automatically incorporate the effects of the excess random variation in their standard errors.

The Logit models proceed in an analogous fashion, but now the link function is the log-odds function:

$$Dead_{ij} \sim Binomial(BatchSize_{ij}, p_i)$$
$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_i \log\left(D_{ij}\right)$$

Here the slope and intercept describe the curve on the log-odds scale. While the direct interpretation of the slope and intercept differ between the Probit and Logit models, the two curves are very similar in shape and will have similar model fits and will give rise to similar estimates of LCxx. The key differences are in the tails of the models where the probit model approaches 0 or 1 more quickly than the logit model.

Unlike the Probit model, the LCxx values can be solved for directly as there is a closed-form analytical solution, i.e.

$$\log\left(\frac{\hat{L}Cxx/100}{1-\hat{L}Cxx/100}\right)=\hat{\beta}_0+\hat{\beta}_i\log(D)$$

gives

$$\log(D)=\frac{\log\left(\frac{\hat{L}Cxx/100}{1-\hat{L}Cxx/100}\right)-\hat{\beta}_0}{\hat{\beta}_i}$$

Similar logit models can be defined with Natural Responses (i.e. non-zero responses in the controls) and random effect models. Goodness-of-fit and corrections for overdispersion are implemented in similar ways.

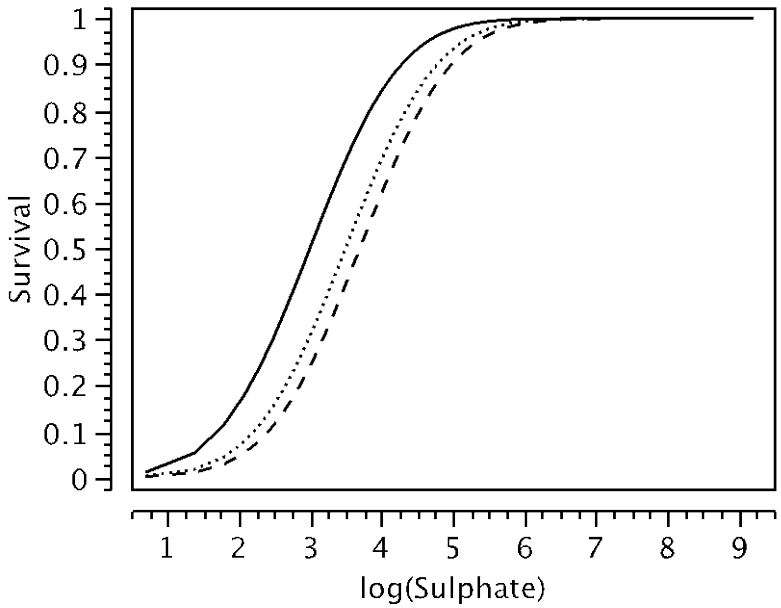
The primary goal of this paper is to investigate the effect of hardness levels on the dose-response curve. We accomplish this by fitting two (or more) models to the combined data from the three hardness levels. In first model (the **Separate** response model), a separate probit/logit curve is fit to each hardness level. So, if the basic probit/logit model is used with 3 hardness levels, this model will require 6 parameters (an intercept and a slope for each hardness level). This can be done in a single model fit rather than (the equivalent) running three separate models (one for each dose). In the second model (the **Common response**) model, the data are pooled over all hardness levels and single probit/logit model is fit. This model has 2 parameters.

The Separate response model is very general. Each hardness level has its own dose-response curve and these curves do not have to have the same shape. Consequently, it is possible that the dose-response curve for lower hardness levels give rise to higher estimated mortalities than the dose-response curve for a higher hardness value. An intermediate model between that of the separate curves for each hardness level and a common curve for all hardness levels, is the **Monotonic-Separate** response curve where the probit/logit curves are "parallel" at different hardness levels and increasing hardness is always "protective", i.e. higher hardness values does not lead to an increase in mortality at any sulphate dose. More formally,

$$Dead_{ij} \sim Binomial(BatchSize_{ij}, p_i)$$

$$p_i = NR + (1-NR)\Phi\left(\beta_0+\beta_i\log\left(D_{ij}\right)+\theta_{hardness}\right)$$

where $\theta_{hardness}$ is the "shift" in the curve due to different hardness's and constraints are placed on these parameters to ensure that the dose-response curve never decreases as hardness increases. [A similar model can be defined for the logit model by changing the link-function.] A schematic of the results from such a model is:



Notice that as hardness increases (the three lines from left to right), the mortality decreases at any sulphate level, i.e. hardness is "protective". The shape of the three curves is identical – all that happens is that the S-shape is shifted to the right as hardness increases.

The suite of potential probit/logit models fit is described by a 3 part "code". First, is the modeling of the effects of hardness as either a separate model for each hardness (**Separate**), or a common model for each hardness (**Common)**, or a model with a shifted-to-the-right ("protective") dose-response curves as hardness increases (**SeparateMono**). Next, does the model assume no natural response (**NoNR**), a common natural response over all hardness levels (**CNR**), or a separate natural response for each hardness level (**SNR**). Finally, does the model include a random effect (**RE**) or exclude random effects (**NoRE**) to account for overdispersion. For example, a probit/logit model identified as **Common, NoNR, NoRE** corresponds to fitting the model with a common curve across all hardness levels, no natural responses, and no random effects.

Other possible choices for quantal responses are Gompertz, log-logistic, etc, but these lead to very similar dose-response curves especially for estimating BMDs in the .10 to .90 range (Ritz, 2010) and so were not fit.

Many of the analyses from EC and NA of the individual studies used isotonic regression (see next section) applied to the observed mortalities. This method would be applicable if there is evidence of a structural lack of fit in the Probit/logit model (i.e. the response is not linear on the probit/logit scale) but no large lack of fit was detected in any of the studies. The isotonic model treats a natural response as simply another set of data values. In these cases, the LCxx values from isotonic regression are not directly comparable to those from the maximum likelihood Probit/logit approach with a natural response. In the isotonic method, no natural response is assumed and so the LCxx value includes the natural response in total mortality.

The choice of model to be fit to a particular study depended on a preliminary inspection of the data (see Schwarz, 2011). For mortality studies that are very sparse (few animals on test) only simple models are tenable (i.e. without natural responses or random effects) can be fit as more complex models will fail to fit because of a lack of clear effect.

For example, consider the plots of fitted models for the EC-RT mortality data found in Figures 1a-1f and summarized in Table 2-EC-RT. The most general model, the Probit, Separate, SNR, RE model (Figure 1a) has a separate dose-response curve for each hardness levels along with a separate natural response curve for each hardness level. This model has a 10 parameter (a slope, intercept, natural response for each of 3 hardness levels plus 1 parameter for the variance of the random effects). The dose-response curve for hardness level 50 is to the left of the dose-response curve for hardness 100 which in turn is to the left of the dose-response curve for hardness 250 in the range of doses studied in this experiment. This ordering is NOT enforced by this model and occurred "naturally" as the data is fairly strong. However, the natural responses do not follow this same ordering with the natural response at hardness 250 is between that of hardness 50 and 100. This model is the most flexible and so has the best fit to the data (largest likelihood value of -165.6 and an AICc value of 356.3.

Figure 1b plots the dose-response curves for the Probit, Separate, CNR, RE model where there are three separate dose-response curves for the three hardness levels, but now all three dose-response curves have the same natural response. This model has 8 parameters (slope and intercept for the 3 dose-response curves plus one parameter for the common natural response plus one parameter for the variance of the random effects). This model is less complex than the previous model (fewer parameters), but will fit the data less well (has a lower likelihood value of -165.8). However, the reduction in fit compared to the previous model is .03 which is a small reduction in fit for a reduction by 2 in the number of parameters because the three separate natural responses from the model in Figure 1a seems to be too flexible as the three natural responses are not very different. Consequently, the AICc of 350.7 is smaller than the AICc of the previous model indicating a model with more support from the data.

Figure 1c plots the dose-response curves for the Probit, SeparateMono, CNR, RE model where the three dose-response curves are "parallel" on the probit scale (which leads to S-shaped curves on the mortality scale that are shifted left or right). While the dose-response curves for this model look very similar to those in Figure 1b (each hardness has a separate dose-response curve), the fit is not as good (the likelihood for this model (-165.9) is slightly less than the likelihood for the previous model (-165.8)). However, this model has fewer parameters (6 in total being the slope and intercept for the first curve plus the variance of the random effects plus the common natural response plus the two shift for hardness levels 100 and 250). Consequently, the AICc is much improved (345.5 for this model vs. 350.7 for the previous model) as the loss in fit (difference in likelihood) is inconsequential relative to the reduction in complexity. This model has better support from the data than the previous two models.

Figures 1d, 1e, and 1f fit the same models as in Figures 1a, 1b and 1c except on the logit scale. The fit is very similar between the probit and logit models with only minor difference in the AIC (Table 2)

Finally, Figure 1g and 1h fits the Probit/Logit, Common, CNR, RE models where a single dose-response curve is fit for all hardness levels. This model has only 4 parameters but has the worst fit to the data (smallest likelihood of -168.1) but the reduction in fit is again offset by the large reduction in the number of parameters required for the fit. The AICc indicates that the Probit model of this form has the most support of all of the models considered in Table 2-EC-RT, but there is no clear distinction to be made between the Probit and Logit models.

Figure 2a-2c illustrates what happens when the SeparateMono model is fit to data that is not monotonic as the hardness level increases. The fit in shown in Figure 2a shows the apparent mortality at hardness 15 is lower than the mortality at hardness 80. The Probit, Separate, NoRN, NoRe model (and all other models where separate curves are fit) does not enforce "protective" effect of hardness. In Figure 2b, the SeparateMono model is fit which enforces a "protective" effect of hardness. Consequently, a single curve is drawn. In fact, this model reduces to the Probit, Common, NoNR, NoRE model [This will only happen in cases with two hardness levels.] Table 2-NA-TA-mortality shows that the likelihood values for the Common and SeparateMono models is the same (implying an indentical fit), but the SeparateMono model has an extra parameter (the effect of hardness 80 relative to hardness 15 which happens to be estimated at 0) and so has less support from the data. Because the Separate model does not enforce the "protective" effect of hardness, it has more support from the data than either of the two other models. In fact, the model with a common LC50 point seems to have the highest support from the data, but there is still substantial support for other models. The sparsity of the data makes it difficult to distinguish among the various models fit to the data. Again, there is little to distinguish between the Probit and Logit models.

4.2 Continuous responses:

There is no common model suitable for modeling weight, reproduction, frond number, or other non-binomials endpoints. The CETIS software has a wide suite of potential models (e.g. the Gompertz) but in the majority of the cases here, the CETIS software uses a linear interpolation method (ICPIN). This is also known as isotonic regression (Barlow et al, 1972). The basic premise is that the response variable should decline with increasing sulphate levels. However, because of sampling fluctuation, the observed curve may not show the monotonic decline with increasing sulphate levels.

Basically, isotonic regression works from left to right through the data. If the mean response at the next *X* value is higher than the current fitted *Y* value, then the previous data and the new *Y* are pooled, a new mean is computed, and algorithm moves to the next *X* value. This is a "non-parametric" method, but can be shown to be the maximum likelihood approach under monotonicity of the sulphate effect. The *R* function *isoreg()*

can be used to fit these models. The likelihood, assuming that the distribution of data values is normally distributed at a particular dose level, can be found from a transformation of the sum-of-squares of the residuals from the fit.

Estimates of the ICxx values are found by linear interpolation on the log(dose) scale. ICxx responses are measured from the mean response at the lowest observable dose rather than at dose 0. For example, if a study used doses 100, 200, 400, 800, 1600 for sulphate, the baseline response is estimated from the dose 100 mean. Because different starting doses were used for different hardness levels, the baseline response may differ among these studies solely because of different initial doses and not because of hardness effects. [A similar problem occurs with functional curves fit as discussed later in this section.] Standard errors (and confidence limits) for the ICxx values are found using a bootstrap method. Several hundred bootstrap samples were generated with replacement from the observed data. For each bootstrap sample, the isotonic regression model was fit and the estimate of the ICxx value determined. The 2.5^{th} and 97.5^{th} percentile of the bootstrap estimates were used as the 95% confidence intervals for the parameter. Note that it is impossible to estimate any ICxx value that exceeds the largest dose observed in the experiment because there is no information from the data on the shape of the curve after the largest observed dose. In these cases, no estimate is reported. Similarly, in some cases, the isotonic regression line is completely flat and no estimate of the ICxx values can be computed.

Isotonic regression models were fit where a single curve was common for all hardness levels (denoted as ***IR.Common)*** or a separate curve was fit for each hardness level (***IR.Separate)***. It is not possible to fit an isotonic regression model with a separate curve for each hardness level but a common ICxx value. I am also unaware of any method that could be used to enforce (declining) monotonicity in the effects of increasing sulphate levels, and increasing monotonicity (i.e. protective effects) in the effects of increasing hardness. Consequently, neither of these two classes of models were fit using isotonic regression.

A second model used for continuous responses in this report is the 3-parameter log-logistic model[4]

$$E[Y] = \frac{A}{\left(1+\left(\frac{X}{D}\right)^{C}\right)}$$

where A is the mean response at X(Dose)=0, D is the dose corresponding to the IC50, and C is a scaling factor. This model was fit using maximum likelihood assuming normally distributed residuals about the fit. The log-likelihood is proportional to the residual sum-of-squares as in the isotonic regression case. Goodness-of-fit was assessed

[4] This model has an additional parameter, σ, representing the standard deviation of the residuals around the fit. This parameter has been included in the AIC computations.

using residual and other diagnostic plots. The ICxx values were found by solving the equation

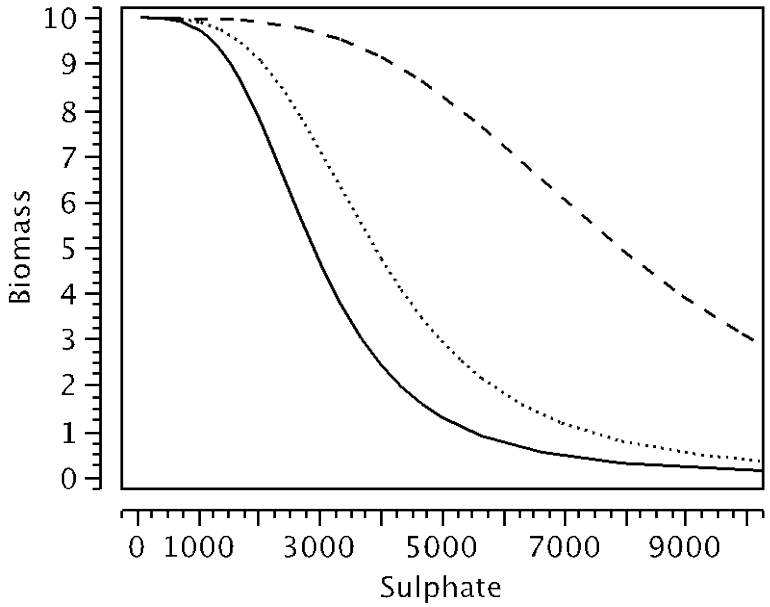
$$A(1-\text{Endpoint}) = \frac{A}{\left(1+\left(\frac{X}{D}\right)^{C}\right)}$$

where the *Endpoint* is .1, .25 or .50. [The above equation can be solved explicitly for *X*.] For example, the IC10 is the dose where the mean response is 90% of the baseline response. It is again possible with sparse data that no estimates of ICxx are available or lie far outside the range of the observed doses in the study. Again, in these cases, the value of ICxx is ignored.

This model can be extended to allow for monotonic (protective) effects of hardness, by adding parameters ($\theta_{hardness}$) which shift the LC50 (the *D*) parameter to the right:

$$E[Y] = \frac{A}{\left(1+\left(\frac{X}{D+\theta_{hardness}}\right)^{C}\right)}$$

For example, the following schematic plot shows three dose-response curves at different hardness levels:



where the hardness effect simply shifts the LC50 to the right which pulls the entire curve to the right compared to the curve at a lower hardness level. This model also assumes a common baseline (response at Dose=0) where as a separate baseline is allowed when a separate curve is fit for each hardness level.

Note that the CETIS software uses the fitted response at the lowest observed dose as the baseline value. For example, if a study used doses 100, 200, 400, 800, 1600 for sulphate,

the baseline response would be measured at dose 100 rather than at dose 0. [Refer to the next model for a worked example.]

Standard errors (and confidence limits) are found using a bootstrap method. Several hundred bootstrap samples were generated with replacement from the observed data. For each bootstrap sample, the log-logistic regression model was fit and the estimate of the ICxx value determined. The 2.5^{th} and 97.5^{th} percentile of the bootstrap estimates were used as the 95% confidence intervals for the parameter.

Three log-logistic models were fit. The **LL3p.Common** model assumed a common curve over all hardness levels; the **LL3p,Separate** model assumed a separate curve for each hardness level, and the **LL3p.Mono** model assumes a shift in the curves to the right with increasing hardness levels.

A third model was used for some responses where there was evidence of an increase in response at lower doses, is the 4-parameter logistic model with hormesis[5]. This models allows the response to increase from baseline before declining as function of dose:

$$E[Y] = \frac{A(1 + E \bullet X)}{(1 + e^{-C(X-D)})}$$

where *A* is a parameter related to baseline value, *D* is the inflection point on the curve (but no longer corresponds to the IC50 point), *C* is a scaling parameter, and *E* is a parameter relating to the hormesis (the increase in response at low doses). The model is fit using maximum likelihood assuming normally distributed errors around the fitted curve. The log-likelihood is then proportional to the residual sum of squares, as in the isotonic regression case. Goodness-of-fit was assessed using residual and other diagnostic plots. The ICxx values were found by solving the equation

$$A(1 - \text{Endpoint}) = \frac{A(1 + E \bullet X)}{(1 + e^{-C(X-D)})}$$

where the *Endpoint* is .1, .25 or .50. [The above equation cannot be solved explicitly for *X* numerical methods must be used to solve the equation.] It is again possible with sparse data that no estimates of ICxx are available or lie far outside the range of the observed doses in the study. Again, in these cases, the value of ICxx is ignored.

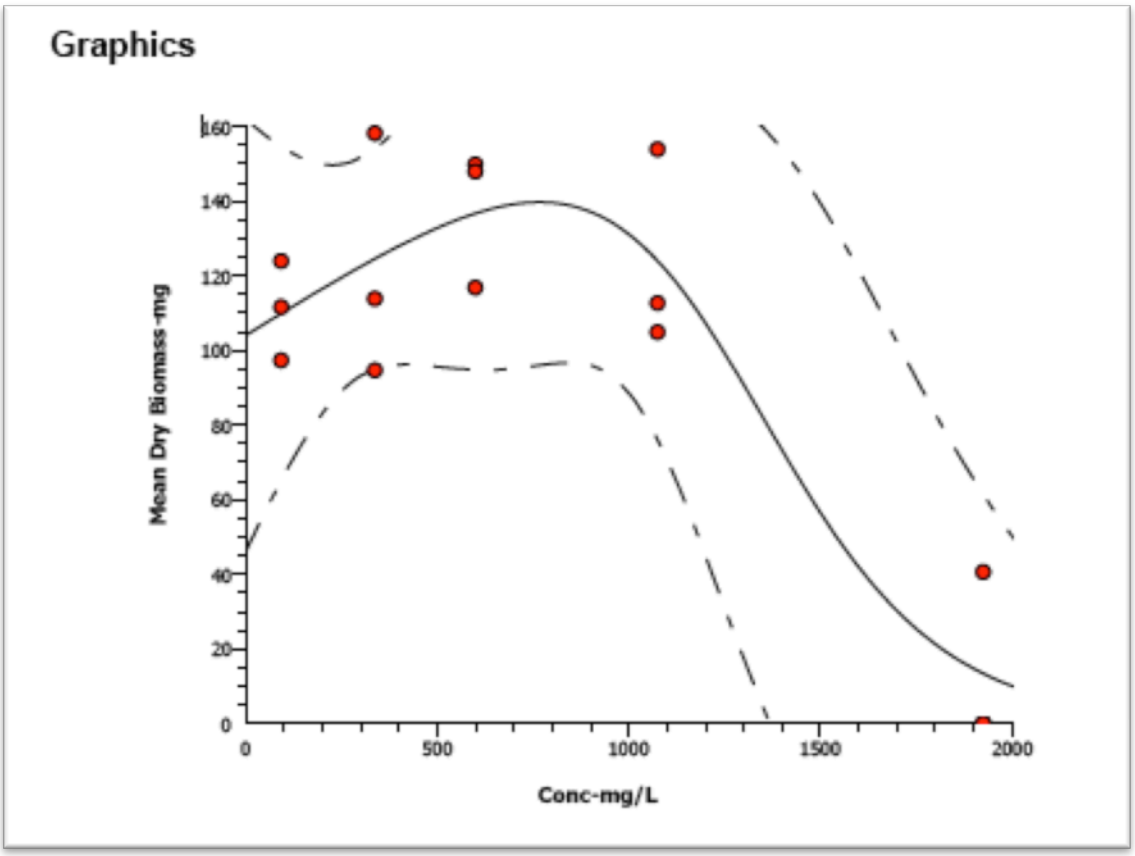
It is not clear how a "protective" effect of hardness should be modeled for this class of models because of the interaction between the parameters in specifying the location of the hump and the decline after the hump. It could be possible to be curves that crossed as sulphate levels increased which would lead to cases where the effect of hardness was protective and not protective depending on the level of sulphate. Consequently, no models that incorporated a monotonic (protective) effect of hardness were fit. Two logistic hormesis models were fit. The **LH4p.Common** model assumed a common curve over all hardness levels; the **LH4p,Separate** model assumed a separate curve for each hardness level.

---

[5] This model has an additional parameter, σ, representing the standard deviation of the residuals around the fit. This parameter has been included in the AIC computations.

As for the log-logistic model, the CETIS software uses the fitted response at the lowest observed dose as the baseline value rather than the estimated response at dose=0. For example, if a study used doses 100, 200, 400, 800, 1600 for sulphate, the baseline response would be measured at dose 100 rather than at dose 0.

For example, consider a graph of the model fit to the Tadpole mean dry biomass in medium hard water (80 mg/L).



The function fit is

$$Y = \frac{A(1 + E \bullet X)}{1 + \exp(-C(X - D))}$$

and the estimated values from the fit were A=104.42, C=-.004325, D=1279, and E=.0006324.

CETIS estimates the IC50 as 1509.9 mg/L. The value of Y at this point is:

$$Y = \frac{A(1 + E \bullet X)}{1 + \exp(-C(X - D))} = \frac{104.42(1 + .0006324 \bullet 1509.9)}{1 + \exp(-(-.004325)(1509.9 - 1279))}$$

$$= 55.06$$

This is NOT 50% of the 104.42 (the value of mean biomass when dose = 0). Rather it is 50% of the mean biomass at dose = 93. At does=93, the response is

$$Y = \frac{A(1+E\bullet X)}{1+\exp(-C(X-D))} = \frac{104.42(1+.0006324\bullet 93)}{1+\exp(-(-.004325)(93-1279))}$$
$$= 110.12$$

For both the 3-parameter log-logistic and the 4-parameter logistic hormesis model, the baseline values for this report were taken as the expected response at dose=0. Consequently, estimates of BMDs may differ in the report from those reported by CETIS.

5. Results.

The model selection table for each species/response listed in Table 1 is presented in Table 2. Plots of the fit of the models to the raw data are available in Schwarz (2011) and a separate document.

For the majority of responses, the majority of model weight was given to models where the dose-response curve is different at the hardness levels tested in the experiments. The results parallels that of Schwarz (2011). This result implies that each hardness level will require a separate experiment to estimate the dose-response curve as there is no “sharing of information” across the hardness levels.

As expected, there was often little to differential between the logit and probit models of the same type.

In the case of Fathead Minnows, the support was split between a model where the dose-response curves were different across hardness levels and a model where the dose-response curves have the same general shape, is parallel at the different hardness levels, and the dose-response curve shifted monotonically as hardness increased. The monotonic-shift model implies that once the dose-response curve is established for any hardness level, the same general curve can be used for higher hardness levels with a simply shift to the right as hardness increases.

In two cases (Lemna frond growth and final weight) there was substantial support for models where the dose-response curve was invariant across hardness levels. In this case, the data are insufficient to distinguish between dose-response curves at different hardness levels. Natural variation in the data and a limited response to sulphate made it difficult to determine the effect of hardness on the dose-response curve.

Finally, in one case (Rainbow Trout mortality) support was almost event split between a model where the dose-response curve was invariant to hardness and where the dose-response curve was parallel and monotonically shifted to the right as hardness increased.

In this case, extra-binomial variation was detected which made it difficult to distinguish the effect of hardness on the dose-response curve.

The general conclusion is that there is strong evidence that the dose-response curve varies by hardness in a non-monotonic fashion for the majority of cases tested in this report. Only in those experiments where there was only a very limited response to sulphate (Lemna frond growth and final weight) or where there was evidence of excess natural variation (Rainbow Trout mortality) was there any substantial support for models where hardness appeared to have no effect on the dose-response curve.

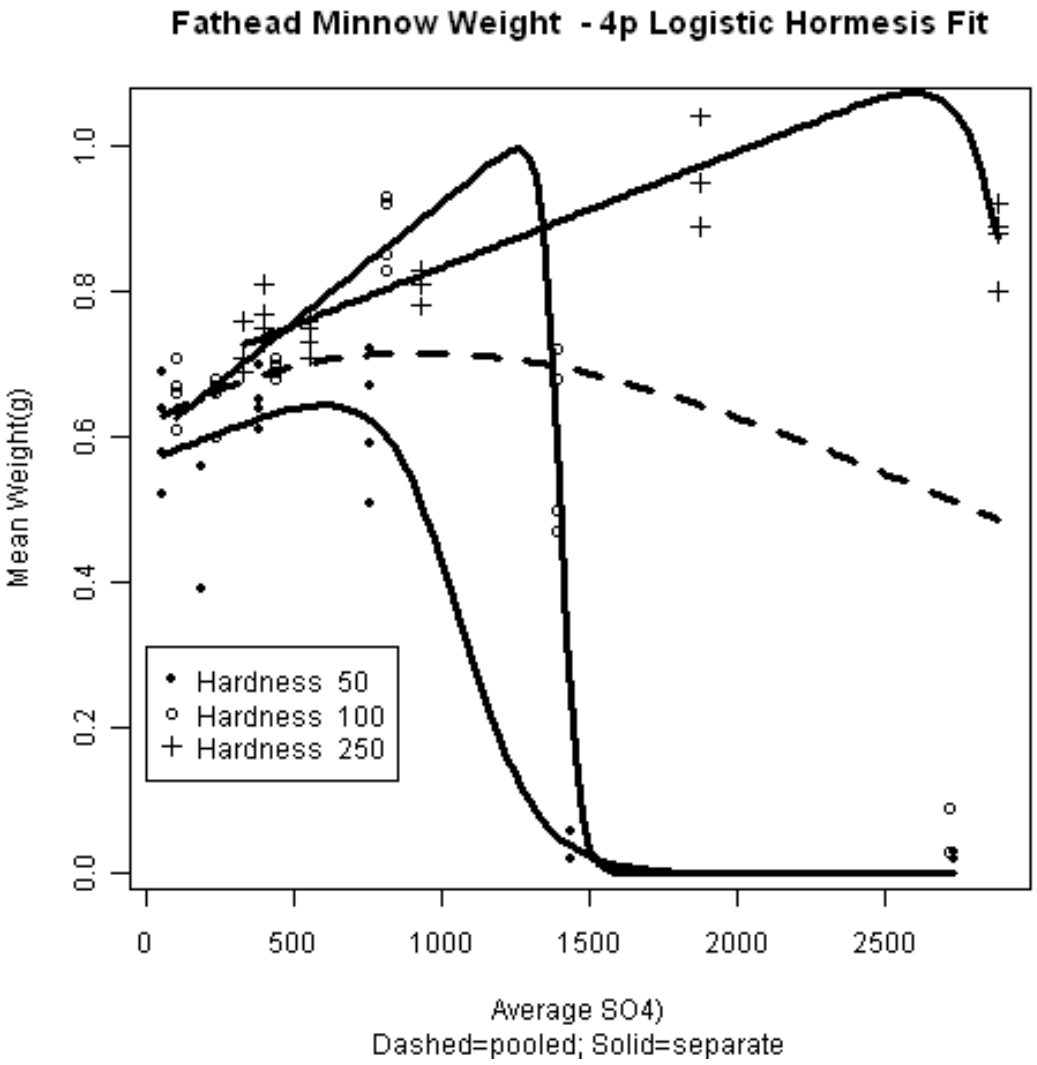
A summary of the model averaged estimates of the LCxx/ICxx are presented in Table 3 with complete details of the individual estimates from each model for each study available as a separate document. For example, Table 4 presents an extract of the model averaging for the LC10 value for EC-Rainbow Trout at hardness 50. Both the common curve over all hardness levels and the separate curve with a monotonic (protective) effect of hardness have substantial support, with minor support for the other models. Estimates of the LC10 (on the log-scale) range from 4.56 (95 on the anti-log scale) to 5.03 (153 on the antilog scale). The weighted average LC10 is 4.81 (on the log scale) corresponding to 123 on the anti-log scale as reported in Table 3. The model average SE incorporates the variability in the estimates among the models fit to the data.

Because the two sets of the top probit/logit pairs models are “contradictory” (one has no effect of hardness while the other has a protective effect of hardness), the model averaged estimates of the LC10 at the three hardness levels (127, 163, and 213 at hardness 50, 100, 250 respectively) are not the same, but are closer together than the estimates from the Separate.Mono model alone (99, 181, and 257 for hardness 50, 100 and 250 respectively as extracted from the Appendix). The model averaged standard errors are larger than the standard errors for any model to account for this model uncertainty.

In some cases, no estimates of the BMD are provided (e.g. estimates of LCxx values at hardness levels 50 and 100 for the mortality studies of EC-Chinook eggs). Examination of the actual data shows that observed mortality was so low, that no model was able to provide sensible estimates of the LCxx values at lower hardness levels.

In some cases, model averaged estimates have very large standard errors. For example, the model averaged LC10 for EC-Fathead Minnows mortality at hardness 250 is 3200 mg/L with a SE=16000! In this case, the observed mortality in the best fitting model at the highest dose was very small, based on only a few organisms, and the extrapolation is not very reliable.

Conversely, the observed standard errors may appear to be very small (e.g. estimate of LC10 for EC-Fathead Minnows mortality at hardness 100 is 1400 (SE 7). In this case, the best fitting model is the 4-parameter logistic hormesis model (see fit below)



The observed data has such a steep decline from increasing in doses up to 1000, then to 0 in higher doses that the curve fit must be very sharp.

Interpretation of the results then follows a two-step process. First, examine the model selection tables (Table 2) to examine the support for models with a common dose-response curve across all hardness levels vs. models with a separate dose-response curve by hardness level. The majority of these tables indicate that there is strong evidence that the dose-response curve varies by hardness. Next consider the model averaged estimated LCxx/ICxx values for each hardness level to see if they differ enough to be biologically important. If the difference in the LCxx values are small across the hardness ranges, then a common LCxx value might be entertained even if the separate model is selected.

6. Discussion

As outlined by Wheeler and Bailer (2005), model averaging provides a way to incorporate model uncertainty into the risk assessment process. Simply selecting the single "best" model may give a false sense of precision (i.e. single model reported standard errors typically underreport the true uncertainty in the BMD).

Model averaging is not a panacea. Estimates of BMDs within the observed range of the doses in a study will typically be very similar across a wide range of models as all of the models must come "close" to the observed data. However, extrapolations that are far outside the observed dose ranges of the data will typically be very sensitive to the choice of models.

It would be possible to extend the above modeling approach by incorporating both the effects of hardness and sulphate upon the observed responses and deriving a single dose-response curve that incorporates both hardness and sulphate levels. The advantage of this more complex approach is that a (model averaged) prediction equation for the BMD can be established as a function of any hardness rather than relying on the observed hardnesses in the study. Unfortunately, in most cases, only a few levels of hardness were studies and so the models for the effect of hardness must be very simple (e.g. linear) and extrapolations outside the observed ranges of hardness will be unwise.

References

Anderson, D. (2008) Model Based Inference in the Life Sciences: A Primer on Evidence. Springer: New York.

Bailer AJ, Wheeler M, Dankovick D, Noble R, Bena J (2005) Incorporating uncertainty and variability in the assessment of occupational hazards. Int J Risk Assess Manage 5:344–357.

Bailer, Noble, and Wheeler (2005) Model uncertainty and risk estimation for experimental studies of quantal responses. Risk Analysis, 25, 291-299.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972) Statistical inference under order restrictions; Wiley, London.

Berkson J (1944). Application of the logistic function to bio-assay (1944). Journal of the American Statistical Association 39: 357–65. doi:10.2307/2280041. JSTOR 2280041

Bliss C.I. (1934). The method of probits. Science 79 (2037): 38–39.

Box, G.E.P. and Draper, N.R. (1987). Empirical Model-Building and Response Surfaces, p. 424, Wiley.

Buday, C. and Schroeder, G. (2011). Biological assessment of sulphate at three water hardness values using various freshwater chronic toxicity tests. Report prepared for the B.C. Ministry of Environment, dated January 2011.

Burnham K.P. and Anderson D.R. (2002) Model selection and multi-model inference: a practical information-theoretic approach. Springer, New York.

Elphick, J. Davies, M., Gilron, G., Canaria, E.C., Lo, B. and Bailey, H. C. (2010). An aquatic toxicological evaluation of sulphate: the case for considering hardness as a modifying factor in setting water quality guidelines. Environmental Toxicology and Chemistry. Online early. DOI: 10.1002/etc.363.

Gibbons. R. D. and Hedeker. D. (1994). Application of random effects probit regression models. Journal of Clincal and Consulting Psychology 62, 285-296.

Gibbons. R. D. Hedeker. D.; Charles. S. C. and Frisch. P. (1994). A random effects probit model for predicting medical malpractice claims. Journal of the American Statistical Association 89, 760-767.

Jeske, D.R. Xu, H.K., Blessinger, T., Jensen, P. Trumble, J. (2009) testing for the Equality of EC50 Values in the Presence of Unequal Slopes With Application to Toxicity of Selenium Types. Journal of Agricultural, Biological and Environmental Statistics 14, 469-483.

Ritz, C. (2010). Toward a unified approach to dose-response modeling in ecotoxicology. Environmental Toxicology and Chemistry, 20, 220-229.

Schwarz, C. J. (2011) A statistical examination of the effect of water hardness on the dose-response of fresh water aquatic species to sulphate. Report prepared for the BC Ministry of Environment, dated 2011-03-28.

Wheeler, M.W. and Bailer, A.J. (2007). Properties of Model-Averaged BMDL's: A study of Model averaging in Dichotomous Response Risk Estimation. Risk Analysis, 27, 659-670.

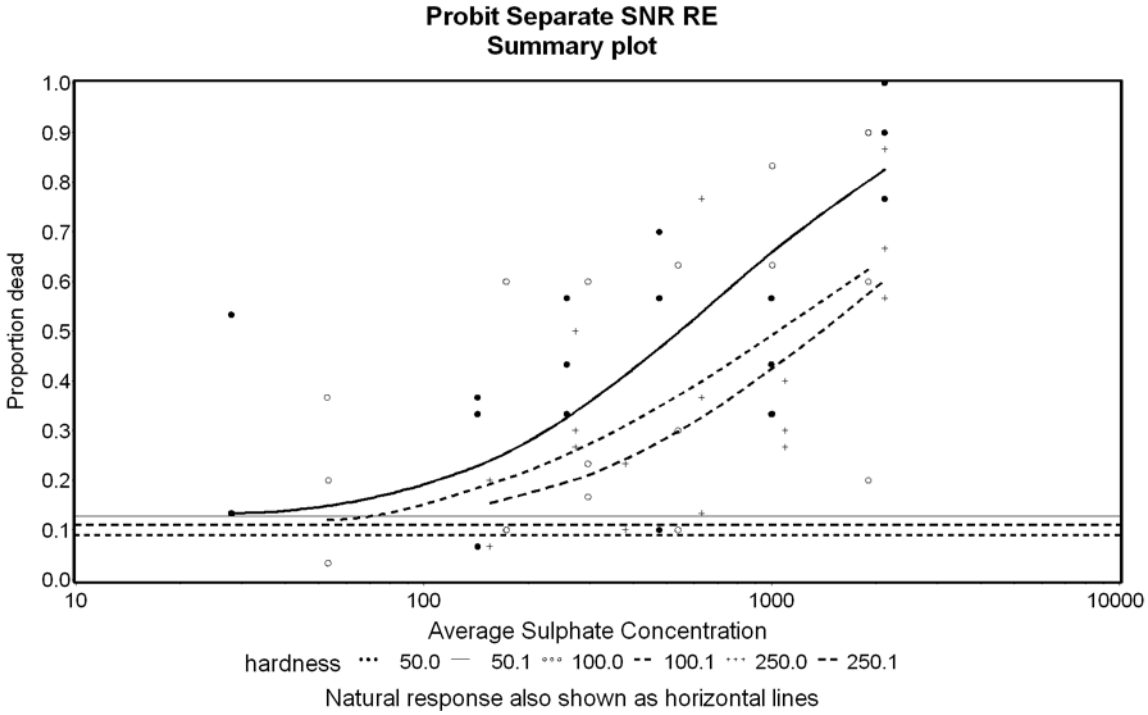


Figure 1a. The Probit, Separate, SNR, RE model as fit to the EC-RT mortality data.

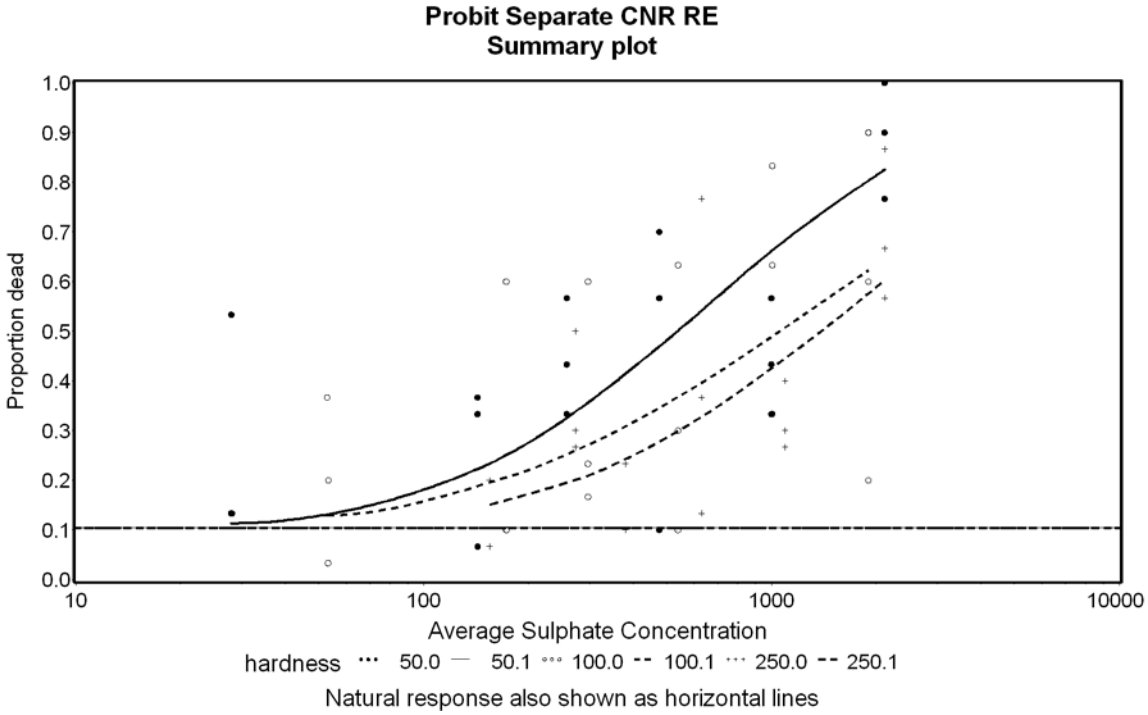


Figure 1b. The Probit, Separate, CNR, RE model as fit to the EC-RT mortality data.

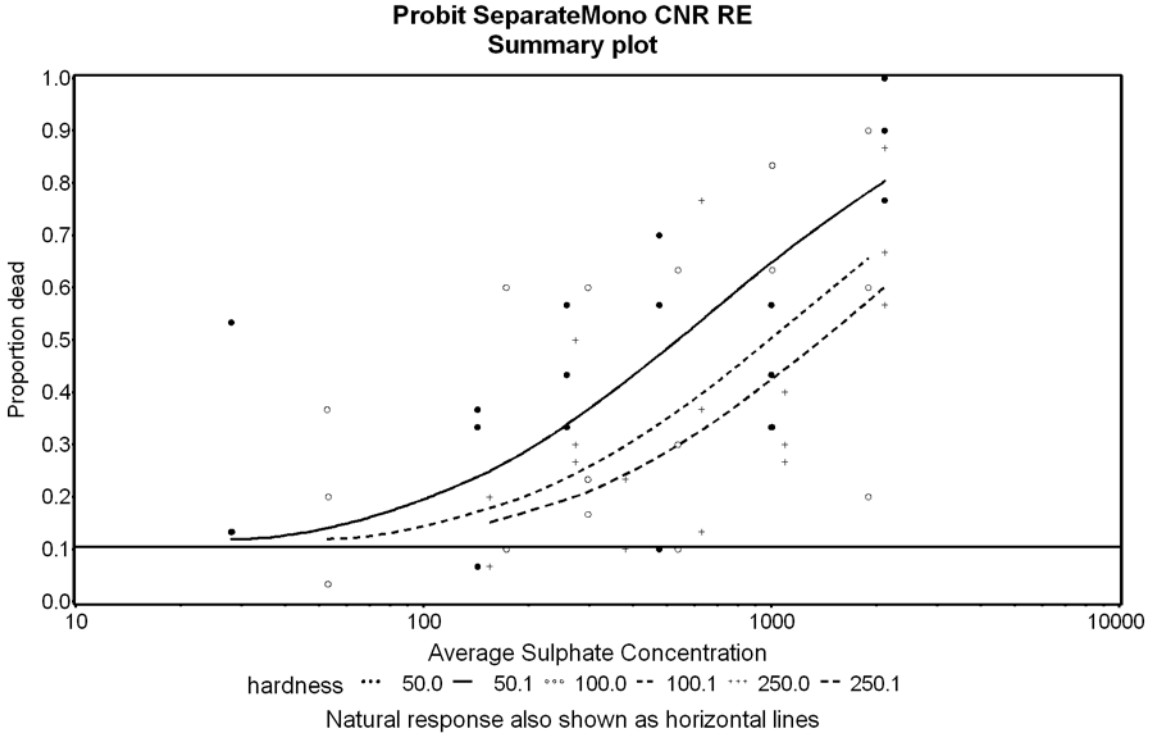


Figure 1c. The Probit, SeparateMono, CNR, RE model as fit to the EC-RT mortality data.

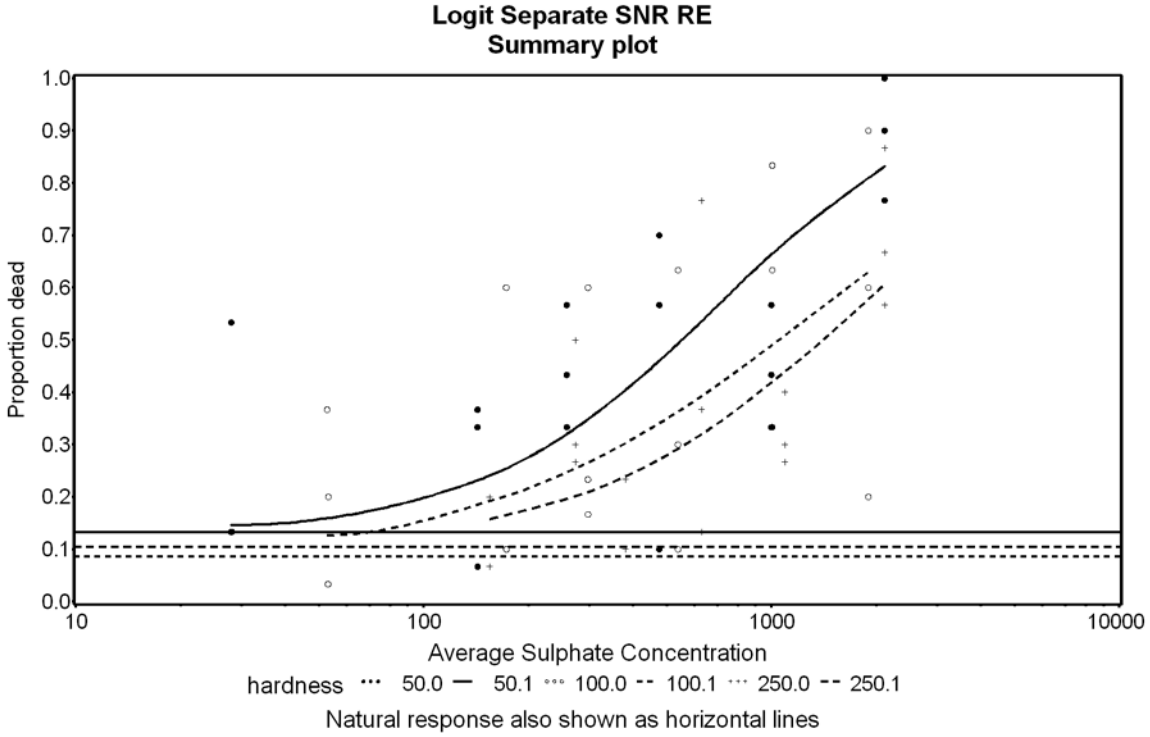


Figure 1d. The Logit, Separate, SNR, RE model as fit to the EC-RT mortality data.

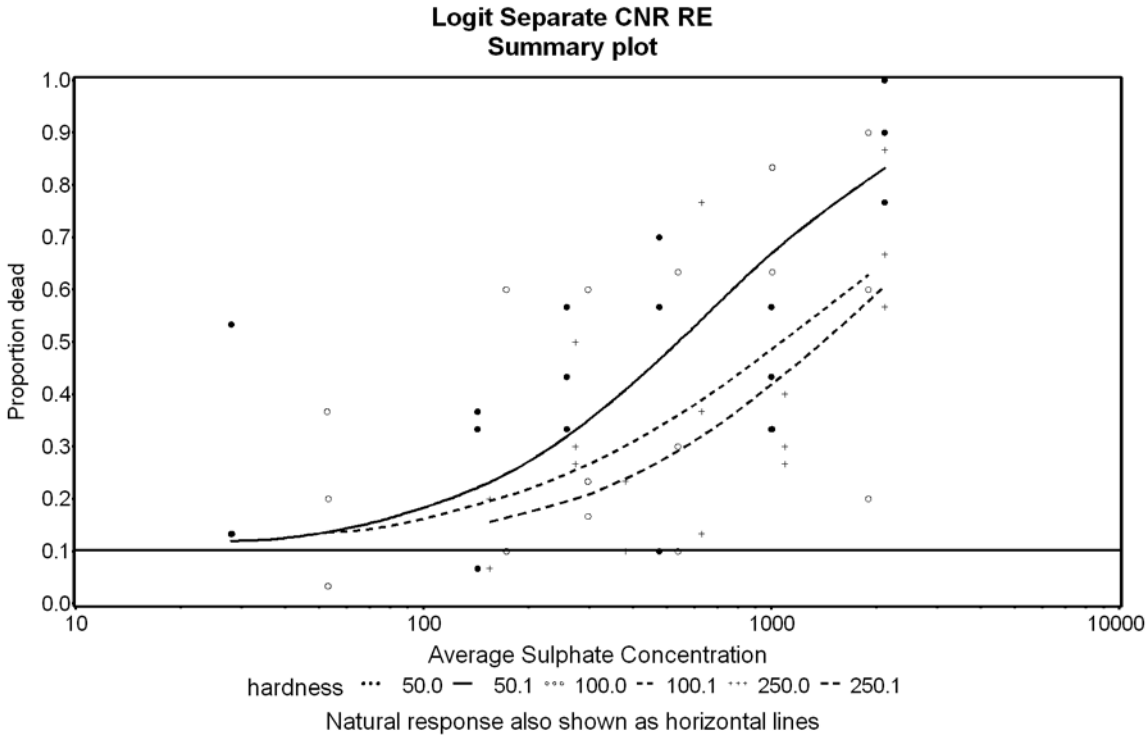


Figure 1e. The Logit, Separate, CNR, RE model as fit to the EC-RT mortality data.

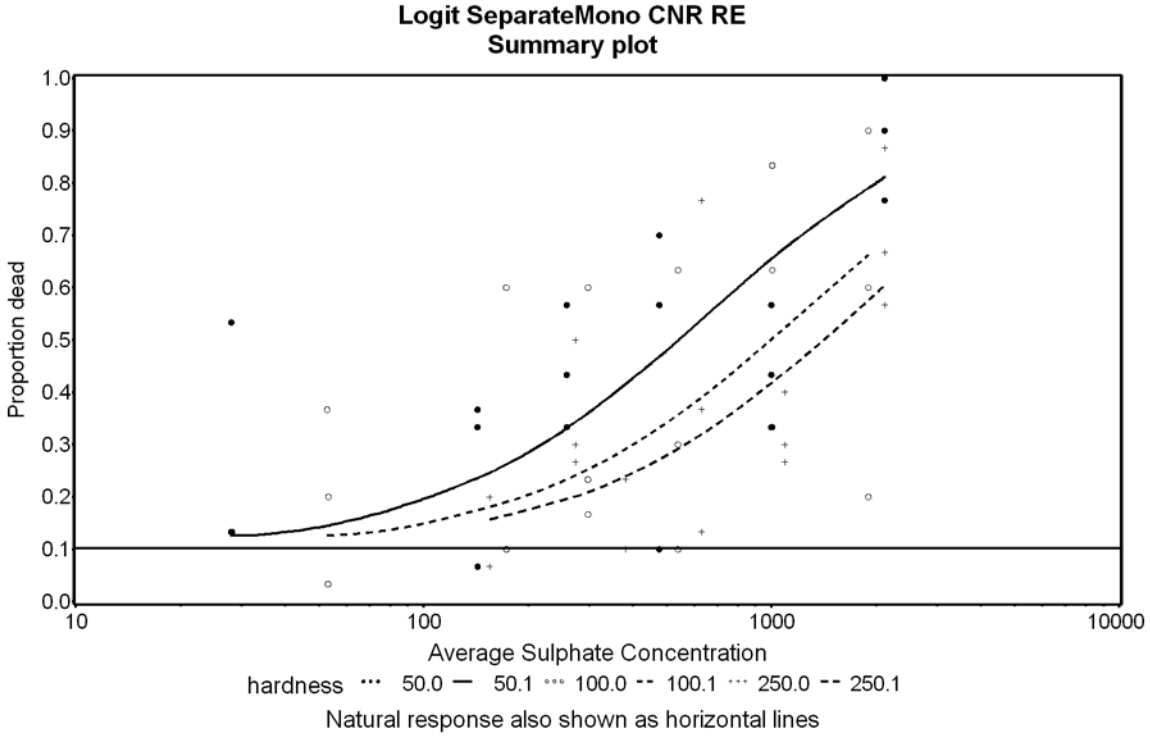


Figure 1f. The Logit, SeparateMono, CNR, RE model as fit to the EC-RT mortality data.

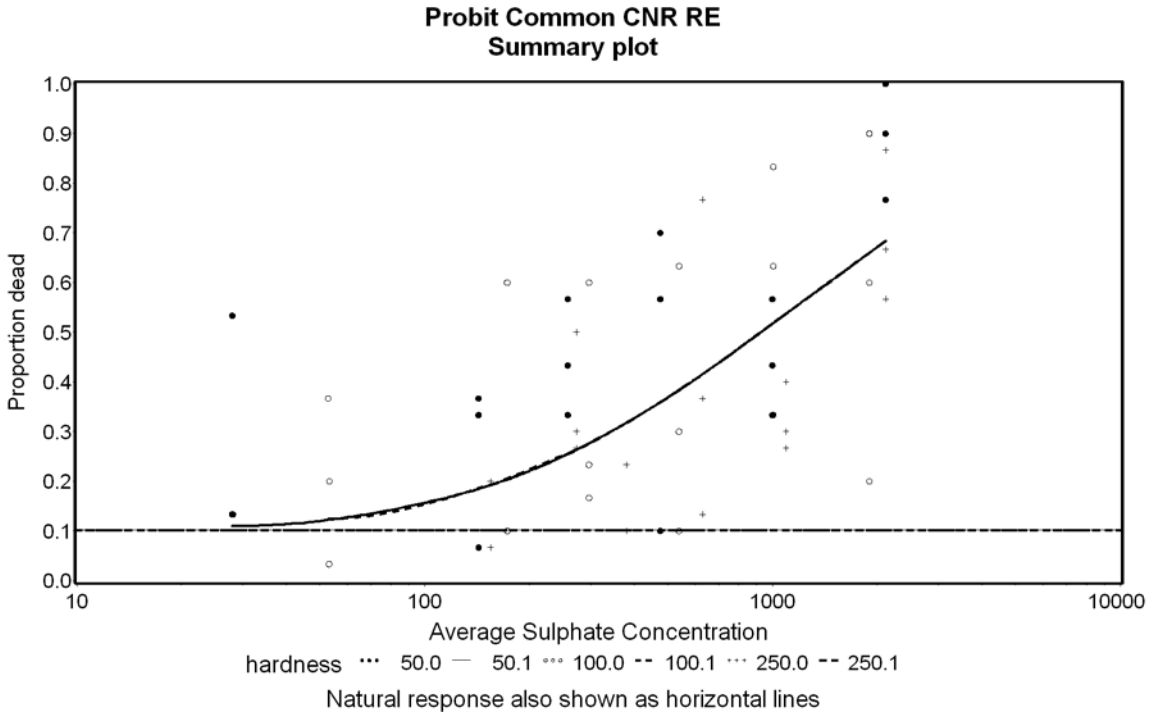


Figure 1g. The Probit, Common, CNR, RE model fit to the EC-RT mortality data.

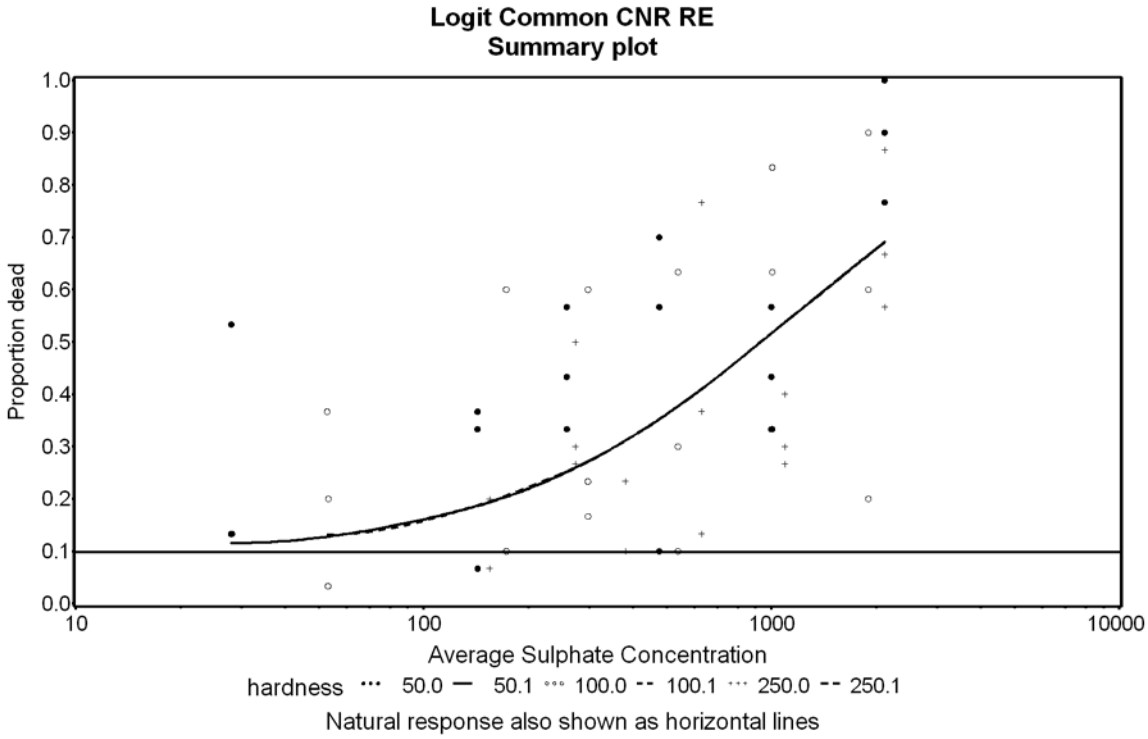


Figure 1h. The Logit, Common, CNR, RE model fit to the EC-RT mortality data.

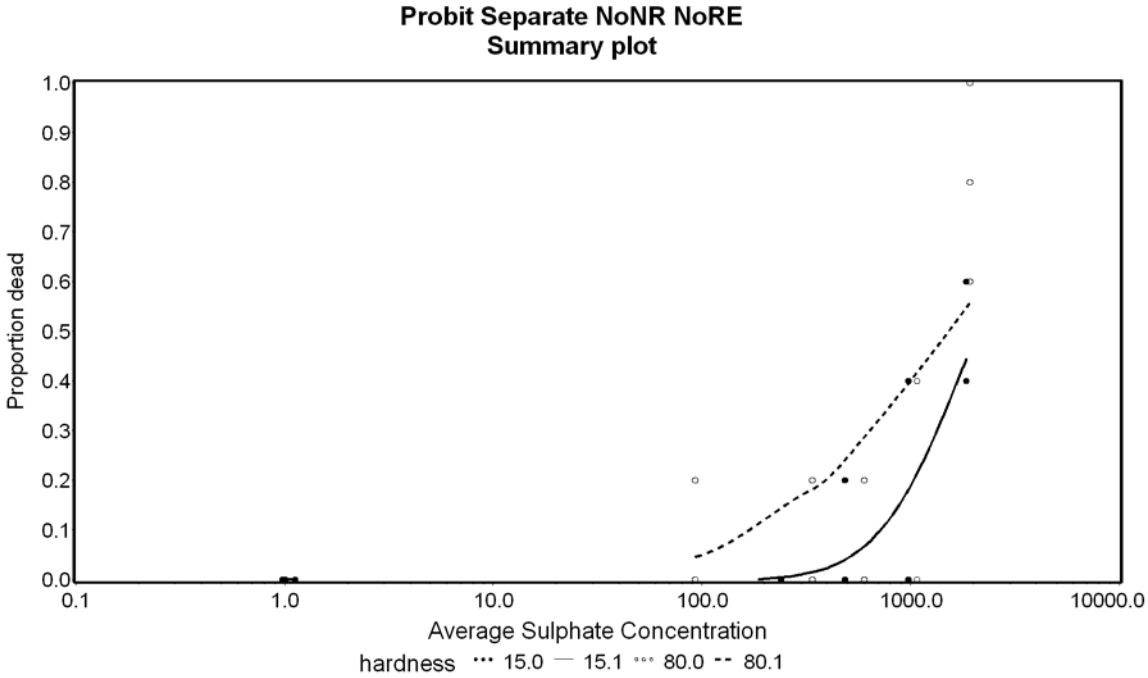


Figure 2a. An illustration of a model fit (Probit, Separate, NoNR, NoRE) where the effect of hardness is not monotonic. This is the fit for the NA-TA-mortality data. The dose response curve at hardness 15 leads to a lower apparent mortality than the same sulphate dose at hardness 80. The Probit, Separate, NoNR, NoRE model does not enforce "protective" effects of hardness.

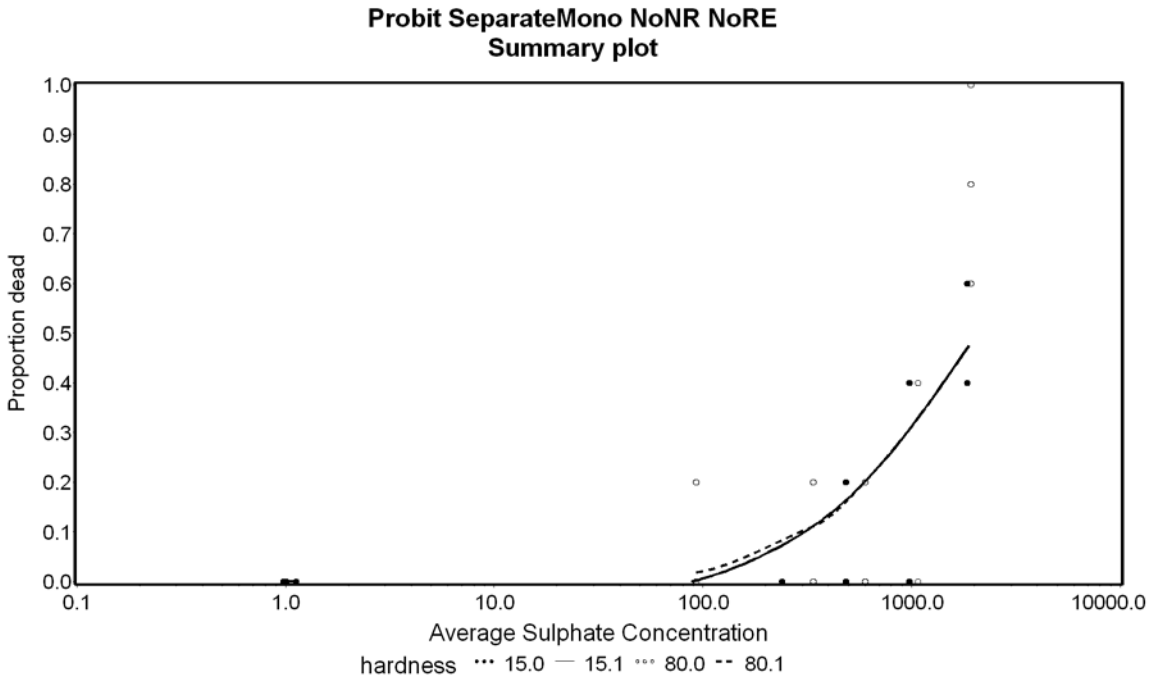


Figure 2b. The Probit, SeparateMono, NoNR, NoRE model fit to the NA-TA-mortality data.

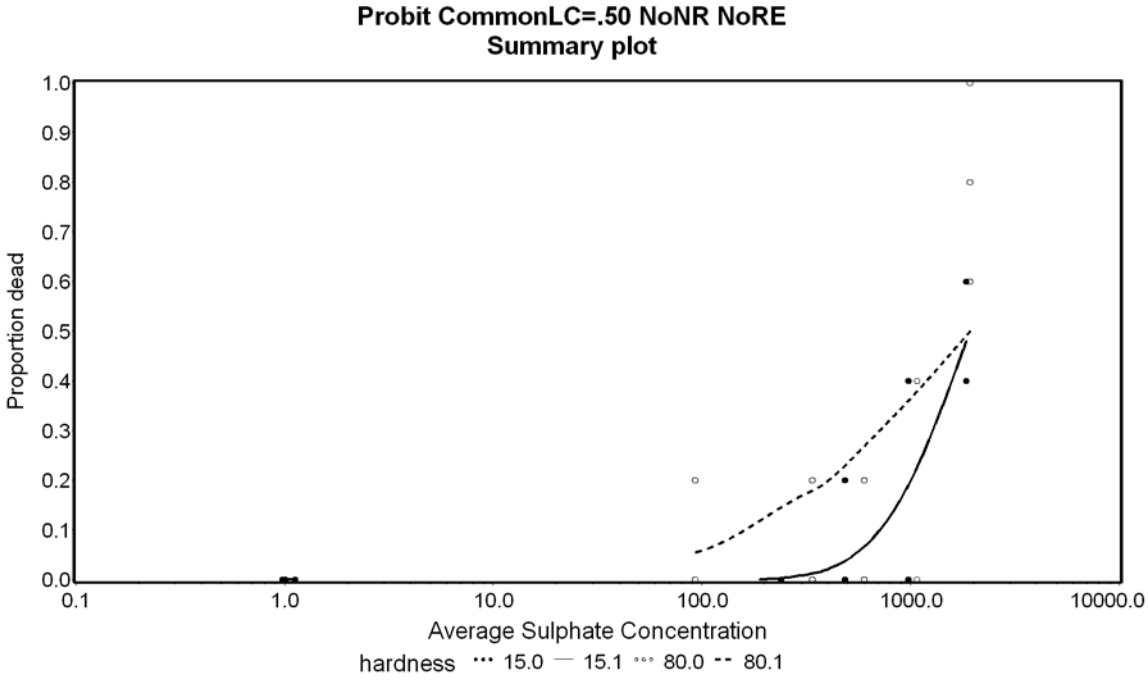


Figure 2c. The Probit, Common LC50, NoNR, NoRe model fit to the NA-TA-mortality data.

Table 1. Summary of sampling protocols for the experiments conducted.		
Environment Canada Studies		
Aquatic species	Response	Sampling protocol at each combination of water hardness and sulphate levels
Rainbow Trout	Survival of eggs to 21 days	Triplicate batches of 30 eggs were incubated and the number of mortalities from each batch was recorded.
Chinook	Survival of eggs to 28 days.	Triplicate batches of 30 eggs were incubated and the number of mortalities from each batch was recorded.
Hyalella	Survival and growth of organisms to 28 days.	Quintuplicate batches (except for 10 batches in the case of control doses of sulphate in soft water) of 15 Hyalella were incubated and the number of mortalities from each batch was recorded. The mean weight of each batch of the organisms at the end of the experiment was measured.
Mussels	Survival and growth of organisms to 28 days.	Triplicate batches of 3, 3, or 4 mussels were incubated and the number of mortalities in each batch was recorded. Wet weight and the beginning and end of the experiment was measured.
Bullfrog tadpoles	Survival and growth to 28 days.	Triplicate batches of 5 tadpoles were incubated and the number of mortalities in each batch was recorded. The change in weight over the 28 days was also recorded.
Fat head minnows	Survival and growth to 7 days.	Quadruplicate batches of 10 minnows were incubated and the number of mortalities in each batch was recorded. The final mean weight in each batch was also recorded.
Lemna	Frond growth and increase in weight	Quadruplicate replicates of Lemna were incubated and the number of new fronds and final weight were recorded for each surviving organism.
Nautilus Studies		
Daphnia	Survival for 6 days and reproduction	10 individual organisms were incubated and the status (dead/alive) and reproductive output was recorded.
Rotifer	Reproduction after 49 hours.	8 individual organisms were incubated and the population growth was recorded.
Fat head minnows	Survival and growth to 7 days.	Triplicate batches of 10 minnows were incubated and the number of mortalities was recorded. The final mean weight in each batch was also recorded.

Bullfrog tadpoles	Survival and growth to 28 days.	Triplicate batches of 5 tadpoles were incubated and the number of mortalities in each batch was recorded. The final biomass was also recorded.
Algae	Cell yield	Four to 10 batches of 10,000 cells were incubated and the percentage increase in the number of cells was recorded.

Table 2-EC-CH-mortality. Summary of AIC model selection for Chinook, mortality conducted by EC

Group	Response	Species	Model Name	# Parameters	Number data values	Log likelihood	AICc	ΔAIC	AICc weight
EC	mortality	CH	Logit, Separate, NoNR, NoRE	6	54	-73.2	160.1	0.0	0.66
EC	mortality	CH	Probit, Separate, NoNR, NoRE	6	54	-73.8	161.5	1.4	0.34
EC	mortality	CH	Logit, Common, NoNR, NoRE	2	54	-91.4	187.1	27.0	0.00
EC	mortality	CH	Probit, Common, NoNR, NoRE	2	54	-91.8	187.9	27.8	0.00
EC	mortality	CH	Logit, SeparateMono, NoNR, NoRE	4	54	-91.4	191.7	31.6	0.00
EC	mortality	CH	Probit, SeparateMono, NoNR, NoRE	4	54	-91.8	192.4	32.3	0.00

Table 2-EC-FH-mortality. Summary of AIC model selection for Fathead Minnow, mortality conducted by EC

Group	Response	Species	Model Name	# Parameters	Number data values	Log likelihood	AICc	ΔAIC	AICc weight
EC	mortality	FH	Logit, Separate, NoNR, NoRE	6	72	-72.9	159.1	0.0	0.94
EC	mortality	FH	Logit, SeparateMono, NoNR, NoRE	4	72	-78.0	164.7	5.6	0.06
EC	mortality	FH	Probit, Separate, NoNR, NoRE	6	72	-83.2	179.6	20.5	0.00
EC	mortality	FH	Probit, SeparateMono, NoNR, NoRE	4	72	-88.4	185.4	26.3	0.00
EC	mortality	FH	Logit, Common, NoNR, NoRE	2	72	-189.5	383.2	224.1	0.00
EC	mortality	FH	Probit, Common, NoNR, NoRE	2	72	-195.4	395.0	236.0	0.00

NoNR=no natural response; NoRE=no random effects

Table 2-EC-FH-weight. Summary of AIC model selection for Fathead Minnow growth conducted by EC.

Group	Response	Species	Model Name	# Parameters	Number data values	Log likelihood	AICc	ΔAIC	AICc weight
EC	weight	FH	LH4p.Separate	15	66	48.2	-56.8	0.0	1.00
EC	weight	FH	IR.Separate	11	66	20.7	-14.6	42.2	0.00
EC	weight	FH	LL3p.Separate	12	66	20.2	-10.5	46.3	0.00
EC	weight	FH	LL3p.Mono	6	66	2.8	7.7	64.5	0.00
EC	weight	FH	IR.Common	4	66	-37.8	84.2	141.0	0.00
EC	weight	FH	LL3p.Common	4	66	-39.9	88.5	145.3	0.00
EC	weight	FH	LH4p.Common	5	66	-39.3	89.6	146.4	0.00

LH4p=4-parameter logistic hormesis models; LL3p = 3-parameter log-logistic model; IR=isotonic regression model.

Table EC-HY-mortality. Summary of AIC model selection for Hyalella, mortality conducted by EC

Group	Response	Species	Model Name	# Parameters	Number data values	Log likelihood	AICc	ΔAIC	AICc weight
EC	mortality	HY	Probit, Separate*, SNR, NoRE	7	100	-142.4	300.0	0.0	0.55
EC	mortality	HY	Logit, Separate*, SNR, NoRE	7	100	-142.6	300.4	0.4	0.45
EC	mortality	HY	Probit, Common, CNR, NoRE	3	100	-153.9	314.0	14.0	0.00
EC	mortality	HY	Logit, Common, CNR, NoRE	3	100	-156.6	319.5	19.5	0.00
EC	mortality	HY	Probit, SeparateMono, CNR, NoRE	5	100	-160.2	331.1	31.1	0.00
EC	mortality	HY	Logit, SeparateMono, CNR, NoRE	5	100	-160.2	331.1	31.1	0.00

CNR=common natural response; SNR=separate natural response; NoRE=no random effects. No dose-response curve as a function of sulphates could be fit for the medium hardness, and so only a natural response was modeled at this hardness.

Table 2-EC-HY-weight. Summary of AIC model selection for Hyalella weight conducted byEC.

Group	Response	Species	Model Name	# Parameters	Number data values	Log likelihood	AICc	ΔAIC	AICc weight
EC	weight	HY	IR.Separate	13	100	33.9	-37.6	0.0	0.69
EC	weight	HY	LL3p.Common	4	100	21.6	-34.7	2.9	0.16
EC	weight	HY	LH4p.Common	5	100	21.6	-32.5	5.1	0.05
EC	weight	HY	IR.Common	7	100	23.6	-32.0	5.6	0.04
EC	weight	HY	LH4p.Separate	15	100	33.6	-31.4	6.2	0.03
EC	weight	HY	LL3p.Mono	6	100	21.6	-30.2	7.4	0.02
EC	weight	HY	LL3p.Separate	12	100	27.6	-27.6	10.0	0.00

LH4p=4-parameter logistic hormesis models; LL3p = 3-parameter log-logistic model; IR=isotonic regression model.

Table 2-EC-LM-frond. Summary of AIC model selection for Lemna, frond growth conducted by EC.

Group	Response	Species	Model Name	# Parameters	Number data values	Log likelihood	AICc	ΔAIC	AICc weight
EC	frond	LM	IR.Common	4	71	-335.0	678.5	0.0	0.62
EC	frond	LM	LH4p.Common	5	71	-334.6	680.1	1.6	0.28
EC	frond	LM	LH4p.Separate	15	71	-322.2	683.1	4.6	0.06
EC	frond	LM	LL3p.Common	4	71	-338.4	685.4	6.9	0.02
EC	frond	LM	LL3p.Mono	6	71	-336.5	686.3	7.8	0.01
EC	frond	LM	IR.Separate	11	71	-333.2	692.9	14.4	0.00
EC	frond	LM	LL3p.Separate	12	71	-335.5	700.4	21.9	0.00

LH4p=4-parameter logistic hormesis models; LL3p = 3-parameter log-logistic model; IR=isotonic regression model.

Table 2-EC-LM-weight. Summary of AIC model selection for Lemna final weight conducted by EC

Group	Response	Species	Model Name	# Parameters	Number data values	Log likelihood	AICc	ΔAIC	AICc weight
EC	weight	LM	IR.Common	3	71	-190.7	387.8	0.0	0.66
EC	weight	LM	LL3p.Common	4	71	-190.8	390.1	2.3	0.21
EC	weight	LM	IR.Separate	6	71	-189.1	391.4	3.6	0.11
EC	weight	LM	LL3p.Mono	6	71	-190.8	394.8	7.0	0.02
EC	weight	LM	LL3p.Separate	12	71	-189.1	407.5	19.7	0.00

LH4p=4-parameter logistic hormesis models; LL3p = 3-parameter log-logistic model; IR=isotonic regression model.

Table 2 EC-MY-mortality. Summary of AIC model selection for Mussels mortality conducted by EC

Group	Response	Species	Model Name	# Parameters	Number data values	Log likelihood	AICc	ΔAIC	AICc weight
EC	mortality	MY	Logit, SeparateMono, NoNR, NoRE	4	54	-44.8	98.5	0.0	0.32
EC	mortality	MY	Probit, SeparateMono, NoNR, NoRE	4	54	-45.3	99.5	1.0	0.20
EC	mortality	MY	Logit, Separate, NoNR, NoRE	6	54	-43.3	100.4	1.9	0.12
EC	mortality	MY	Logit, Common, CNR, NoRE	3	54	-47.0	100.6	2.1	0.12
EC	mortality	MY	Probit, Separate, NoNR, NoRE	6	54	-43.7	101.2	2.7	0.08
EC	mortality	MY	Logit, Common, NoNR, NoRE	2	54	-48.9	101.9	3.4	0.06
EC	mortality	MY	Probit, Common, NoNR, NoRE	2	54	-49.0	102.3	3.8	0.05
EC	mortality	MY	Logit, Separate, SNR, NoRE	9	54	-40.3	102.8	4.3	0.04
EC	mortality	MY	Probit, Separate, SNR, NoRE	9	54	-41.5	105.0	6.5	0.01
EC	mortality	MY	Probit, Common, CNR, NoRE	3	54	-51.8	110.2	11.7	0.00

NoNR=no natural response; NoRE=no random effects

Table 2-EC-RT-mortality. Summary of AIC model selection for Rainbow Trout mortality conducted by EC.

Group	Response	Species	Model NameS	# Parameters	Number data values	Log likelihood	AICc	ΔAIC	AICc weight
EC	mortality	RT	Probit, Common, CNR, RE	4	54	-168.1	345.1	0.0	0.27
EC	mortality	RT	Logit, Common, CNR, RE	4	54	-168.2	345.2	0.1	0.26
EC	mortality	RT	Probit, SeparateMono, CNR, RE	6	54	-165.9	345.5	0.4	0.22
EC	mortality	RT	Logit, SeparateMono, CNR, RE	6	54	-165.9	345.6	0.5	0.22
EC	mortality	RT	Probit, Separate, CNR, RE	8	54	-165.8	350.7	5.6	0.02
EC	mortality	RT	Logit, Separate, CNR, RE	8	54	-165.8	350.8	5.7	0.02
EC	mortality	RT	Logit, Separate, SNR, RE	10	54	-165.5	356.2	11.1	0.00
EC	mortality	RT	Probit, Separate, SNR, RE	10	54	-165.6	356.3	11.2	0.00

SNR=separate natural response; CNR=common natural response; RE=random effects

Table 2-NA-AL-cell. Summary of AIC model selection for Algae cell increases conducted by NA.

Group	Response	Species	Model Name	# Parameters	Number data values	Log likelihood	AICc	ΔAIC	AICc weight
NA	cell.incre	AL	LL3p.Separate	12	80	-329.8	688.2	0.0	0.96
NA	cell.incre	AL	LL3p.Mono	6	80	-340.8	694.8	6.6	0.04
NA	cell.incre	AL	LL3p.Common	4	80	-353.7	715.9	27.7	0.00
NA	cell.incre	AL	IR.Common	15	80	-346.6	730.6	42.4	0.00
NA	cell.incre	AL	IR.Separate	25	80	-329.4	732.9	44.7	0.00

LH4p=4-parameter logistic hormesis models; LL3p = 3-parameter log-logistic model; IR=isotonic regression model.

Table 2-NA-DA-mortality. Summary of AIC model selection for Daphnia mortality conducted by NA.

Group	Response	Species	Model Name	# Parameters	Number data values	Log likelihood	AICc	ΔAIC	AICc weight
NA	mortality	DA	Logit, SeparateMono, NoNR, NoRE	5	320	-71.0	152.3	0.0	0.62
NA	mortality	DA	Logit, Separate, NoNR, NoRE	8	320	-68.5	153.5	1.2	0.33
NA	mortality	DA	Probit, Separate, NoNR, NoRE	8	320	-71.1	158.6	6.4	0.03
NA	mortality	DA	Probit, SeparateMono, NoNR, NoRE	5	320	-74.9	160.0	7.7	0.01
NA	mortality	DA	Logit, Common, NoNR, NoRE	2	320	-79.2	162.4	10.1	0.00
NA	mortality	DA	Probit, Common, NoNR, NoRE	2	320	-84.1	172.3	20.0	0.00

NoNR=no natural response; NoRE=no random effects

Table 2-NA-DA-repro. Summary of model selection for Daphnia reproduction conducted by NA.

Group	Response	Species	Model Name	# Parameters	Number data values	Log likelihood	AICc	ΔAIC	AICc weight
NA	repro	DA	LL3p.Separate	16	318	-1349.3	2732.3	0.0	1.00
NA	repro	DA	LL3p.Mono	7	318	-1406.5	2827.4	95.1	0.00
NA	repro	DA	LL3p.Common	4	318	-1458.6	2925.3	193.0	0.00
NA	repro	DA	IR.Separate	113	318	-1337.4	3027.2	294.9	0.00
NA	repro	DA	IR.Common	88	318	-1445.4	3135.2	402.9	0.00

LH4p=4-parameter logistic hormesis models; LL3p = 3-parameter log-logistic model; IR=isotonic regression model.

Table 2-NA-FH-mortality. Summary of model selection for Fathead Minnow mortality conducted by NA.

Group	Response	Species	Model Name	# Parameters	Number data values	Log likelihood	AICc	ΔAIC	AICc weight
NA	mortality	FH	Logit, SeparateMono, NoNR, NoRE	5	96	-130.5	271.7	0.0	0.64
NA	mortality	FH	Logit, Separate, NoNR, NoRE	8	96	-127.6	272.9	1.2	0.36
NA	mortality	FH	Probit, Separate, NoNR, NoRE	8	96	-135.9	289.5	17.8	0.00
NA	mortality	FH	Probit, SeparateMono, NoNR, NoRE	5	96	-140.1	290.8	19.1	0.00
NA	mortality	FH	Logit, Common, NoNR, NoRE	2	96	-179.7	363.5	91.8	0.00
NA	mortality	FH	Probit, Common, NoNR, NoRE	2	96	-187.6	379.3	107.6	0.00

NoNR=no natural response; NoRE=no random effects

Table 2-NA-FH-weight. Summary of AIC model selection for Fathead Minnow weight conducted by NA.

Group	Response	Species	Model Name	# Parameters	Number data values	Log likelihood	AICc	ΔAIC	AICc weight
NA	weight	FH	LL3p.Mono	7	91	33.3	-51.2	0.0	0.55
NA	weight	FH	LL3p.Separate	16	91	45.1	-50.8	0.4	0.45
NA	weight	FH	IR.Separate	27	91	52.1	-26.2	25.0	0.00
NA	weight	FH	LL3p.Common	4	91	-1.1	10.6	61.8	0.00
NA	weight	FH	IR.Common	10	91	4.3	14.2	65.4	0.00

LH4p=4-parameter logistic hormesis models; LL3p = 3-parameter log-logistic model; IR=isotonic regression model.

Table 2-NA-RO-reproduction. Summary of AIC model selection for Rotifers reproduction conducted by NA.

Group	Response	Species	Model Name	# Parameters	Number data values	Log likelihood	AICc	ΔAIC	AICc weight
NA	growth	RO	IR.Separate	19	140	-130.1	304.4	0.0	0.79
NA	growth	RO	LL3p.Separate	16	140	-135.8	308.0	3.6	0.13
NA	growth	RO	IR.Common	8	140	-146.5	310.2	5.8	0.04
NA	growth	RO	LL3p.Mono	7	140	-148.1	311.1	6.7	0.03
NA	growth	RO	LL3p.Common	4	140	-152.6	313.4	9.0	0.01

LH4p=4-parameter logistic hormesis models; LL3p = 3-parameter log-logistic model; IR=isotonic regression model.

Table 2-NA-TA-mortality. Summary of model selection for Tadpole mortality conducted by NA.

Group	Response	Species	Model Name	# Parameters	Number data values	Log likelihood	AICc	ΔAIC	AICc weight
NA	mortality	TA	Logit, Separate, NoNR, NoRE	4	30	-29.3	68.3	0.0	0.43
NA	mortality	TA	Probit, Separate, NoNR, NoRE	4	30	-29.9	69.3	1.1	0.25
NA	mortality	TA	Logit, Common, NoNR, NoRE	2	30	-32.8	70.0	1.7	0.18
NA	mortality	TA	Probit, Common, NoNR, NoRE	2	30	-33.9	72.2	3.9	0.06
NA	mortality	TA	Logit, SeparateMono, NoNR, NoRE	3	30	-32.8	72.5	4.2	0.05
NA	mortality	TA	Probit, SeparateMono, NoNR, NoRE	3	30	-33.9	74.6	6.4	0.02

LH4p=4-parameter logistic hormesis models; LL3p = 3-parameter log-logistic model; IR=isotonic regression model.

Table 2-NA-TA-weight. Summary of model selection for Tadpole weight gain conducted by NA.

Group	Response	Species	Model Name	# Parameters	Number data values	Log likelihood	AICc	ΔAIC	AICc weight
NA	weight	TA	IR.Separate	7	30	-140.8	300.7	0.0	0.71
NA	weight	TA	LL3p.Common	4	30	-147.4	304.4	3.7	0.11
NA	weight	TA	LL3p.Separate	8	30	-140.8	304.5	3.8	0.11
NA	weight	TA	LH4p.Common	5	30	-147.4	307.3	6.6	0.03
NA	weight	TA	LL3p.Mono	5	30	-147.4	307.3	6.6	0.03
NA	weight	TA	IR.Common	6	30	-146.5	308.7	8.0	0.01
NA	weight	TA	LH4p.Separate	10	30	-139.3	310.2	9.5	0.01

LH4p=4-parameter logistic hormesis models; LL3p = 3-parameter log-logistic model; IR=isotonic regression model.

Table 3. Summary of model averaged LCxx/ICxx values. Estimates that are far outside the range of the observed doses are not reported.													
				BMD									
				LC/IC 10				LC/IC 25				LC/IC 50			
				MA est	MA se	MA LCL	MA UCL	MA est	MA se	MA LCL	MA UCL	MA est	MA se	MA LCL	MA UCL
Group	Species	Response	Hardness												
EC	CH	mortality	50	.	.	.	.	.	.	.	.	.	.	.	.
			100	.	.	.	.	.	.	.	.	.	.	.	.
			250	1287	147	1028	1610	2521	428	1807	3517	.	.	.	.
	FH	mortality	50	379	56	283	506	598	62	489	732	946	80	801	1116
			100	1120	110	924	1357	1436	101	1252	1649	1843	118	1626	2089
			250	3092	234	2666	3586	3085	86	2921	3259	3178	90	3007	3358
		weight	50	931	159	666	1301	1004	136	771	1308	1111	104	924	1336
			100	1397	7	1383	1411	1408	7	1394	1422	1428	7	1414	1442
			250	2969	12	2946	2992	2999	12	2975	3023	3053	12	3030	3077
	HY	mortality	50	1430	247	1020	2005	2178	284	1687	2812	3404	824	2118	5471
			100	.	.	.	.	.	.	.	.	.	.	.	.
			250	.	.	.	.	.	.	.	.	.	.	.	.
		weight	50	1170	434	566	2420	1739	423	1080	2801	.	.	.	.
			100	682	323	269	1727	1030	271	616	1724	.	.	.	.
			250	437	245	145	1314	1198	369	656	2191	1929	385	1305	2852
	LM	frond	50	2143	3228	112	41052	.	.	.	.	.	.	.	.

Table 3. Summary of model averaged LCxx/ICxx values. Estimates that are far outside the range of the observed doses are not reported.															
				BMD											
				LC/IC 10				LC/IC 25				LC/IC 50			
				MA est	MA se	MA LCL	MA UCL	MA est	MA se	MA LCL	MA UCL	MA est	MA se	MA LCL	MA UCL
			100	2243	3290	127	39740	.	.	.	.	.	.	.	.
			250	2314	3258	147	36545	.	.	.	.	.	.	.	.
		weight	50	.	.	.	.	.	.	.	.	.	.	.	.
			100	.	.	.	.	.	.	.	.	.	.	.	.
			250	.	.	.	.	.	.	.	.	.	.	.	.
	MY	mortality	50	139	175	12	1640	730	569	158	3360	.	.	.	.
			100	.	.	.	.	.	.	.	.	.	.	.	.
			250	676	798	67	6842	.	.	.	.	.	.	.	.
	RT	mortality	50	123	62	45	333	322	126	149	694	889	353	408	1936
			100	162	74	66	395	427	131	233	780	1189	370	645	2189
			250	191	97	71	517	502	190	239	1055	1392	521	668	2898

Table 3. Summary of model averaged LCxx/ICxx values. Estimates that are far outside the range of the observed doses are not reported.

				BMD											
				LC/IC 10				LC/IC 25				LC/IC 50			
				MA est	MA se	MA LCL	MA UCL	MA est	MA se	MA LCL	MA UCL	MA est	MA se	MA LCL	MA UCL
NA	AL	cell.incre	10	441	182	196	988	696	184	415	1168	1101	165	821	1477
			80	2487	100	2300	2690	2615	60	2500	2736	2749	25	2701	2798
			320	2548	43	2464	2634	2660	21	2618	2702	2777	17	2744	2810
	DA	mortality	40	402	75	279	581	570	89	419	775	809	115	612	1071
			80	593	133	382	920	871	145	628	1208	1282	188	962	1708
			160	857	160	594	1237	1145	159	872	1504	1531	198	1189	1972
			320	816	122	609	1095	1135	144	884	1456	1580	198	1236	2019
		repro	40	158	217	11	2331	272	208	61	1215	468	183	217	1009
			80	708	249	356	1409	890	196	578	1369	1119	117	911	1374
			160	1184	9	1166	1203	1223	5	1213	1233	1263	5	1253	1273
			320	253	202	53	1210	425	235	144	1257	717	270	343	1498
	FH	mortality	40	352	68	241	515	743	108	558	988	1565	212	1199	2041
			80	464	91	316	681	1043	151	786	1384	2344	349	1751	3137
			160	1244	240	853	1815	2549	383	1898	3423	5222	955	3649	7472
			320	2516	623	1548	4089	6376	2552	2910	13972	.	.	.	.
		weight	40	600	168	346	1038	869	155	612	1233	1260	162	979	1621

Table 3. Summary of model averaged LCxx/ICxx values. Estimates that are far outside the range of the observed doses are not reported.														
				BMD										
				LC/IC 10				LC/IC 25				LC/IC 50			
				MA est	MA se	MA LCL	MA UCL	MA est	MA se	MA LCL	MA UCL	MA est	MA se	MA LCL	MA UCL
			80	1330	243	930	1904	1845	246	1421	2396	2559	310	2018	3244
			160	2102	561	1246	3548	2809	536	1932	4083	3752	756	2528	5568
			320	.	.	.	.	716	4705	0	2.795E8	4304	2196	1584	11698
	RO	**growth**	**40**	733	1074	41	12955	995	260	597	1660	1211	289	759	1933
			80	352	309	63	1969	1799	603	933	3469	2191	1072	840	5717
			160	724	278	341	1536	1311	1226	209	8201	.	.	.	.
			320	848	677	177	4059	1071	697	299	3837	.	.	.	.
	TA	**mortality**	**15**	587	249	256	1346	1068	275	645	1769	1986	500	1212	3255
			80	242	121	90	646	607	191	328	1124	1583	507	845	2964
		weight	**15**	1246	1397	138	11213	1441	1487	191	10881	1828	44	1744	1917
			80	1276	419	671	2429	1385	337	860	2231	1577	231	1184	2100

Table 4. Example of model averaging for estimates of LC10 for Rainbow Trout at hardness 50 conducted by EC.

Report	Response	Species	Model	Estimate	Standard Error	AICc	Delta AICc	AIC weight	Model Average Estimate	Model Average SE	95% ci lower bound	95% ci upper bound
EC	mortality	RT	Probit, Common, CNR, RE	5.03	0.42	345.1	0.0	0.27	.	.	.	.
EC	mortality	RT	Logit, Common, CNR, RE	4.98	0.43	345.2	0.1	0.26	.	.	.	.
EC	mortality	RT	Probit, SeparateMono, CNR, RE	4.60	0.49	345.5	0.4	0.22	.	.	.	.
EC	mortality	RT	Logit, SeparateMono, CNR, RE	4.56	0.51	345.6	0.5	0.22	.	.	.	.
EC	mortality	RT	Probit, Separate, CNR, RE	4.73	0.55	350.7	5.6	0.02	.	.	.	.
EC	mortality	RT	Logit, Separate, CNR, RE	4.70	0.57	350.8	5.7	0.02	.	.	.	.
EC	mortality	RT	Logit, Separate, SNR, RE	4.85	0.60	356.2	11.1	0.00	.	.	.	.
EC	mortality	RT	Probit, Separate, SNR, RE	4.84	0.58	356.3	11.2	0.00	.	.	.	.
EC	mortality	RT	99-Model Averaged	.	.	.	.	.	4.81	0.51	3.81	5.81
EC	mortality	RT	99-Model Averaged on antilog	.	.	.	.	.	122.73	62.49	45.24	332.94