OPINION PIECE

NECS, NOECS AND THE EC_x

David R Fox
Australian Centre for Environmetrics, University of Melbourne, Parkville, Vic 3010 Australia.

Australian researchers have made a number of significant contributions to the science of ecological toxicity (ecotox) testing. A significant development was the generalisation of the method proposed by Aldenberg and Slob (1993) for setting confidence limits on a hazardous concentration obtained from a species sensitivity distribution (SSD). In 1999 I was involved in the re-write of the ANZECC (1992) Water Quality Guidelines and became interested in statistical issues associated with SSDs. Encouraged by earlier successes with the use of Burr's distributions in the context of power systems simulation (Fox 1978), I suggested to Dr Shao (CSIRO, Mathematical and Information Sciences) that he explore the possibility of using this class of distributions for SSD-related work. As it turned out, the relationship between the Burr Type III and log-logistic distributions has been established almost 20 years earlier by Tadikamalla (1980). Using this result, Shao (2000) suggested the use of the Burr Type III distribution as a more flexible approach to SSD modelling. The methodology was ultimately embodied in a software tool called BurrliOZ developed by CSIRO (http://www.cmis.csiro.au/envir/burrlioz/). The BurrliOZ software was distributed with the ANZECC/ARMCANZ (2000) Water Quality Guidelines and this, perhaps more than any other single factor helped ensure that this highly statistical method became *de facto* practice in Australia and New Zealand for setting risk-based trigger values.

While no system or approach to setting protective environmental concentrations is perfect, the so-called statistical extrapolation method has proven to be reasonably robust. Having said that, after eight years of 'road-testing' the BurrliOZ methodology, it is perhaps time to stand back and ask 'how well is it doing?'. Early concerns were expressed with the potential for BurrliOZ to generate unrealistically low trigger values in some instances. Such anomalies were acknowledged in the ANZECC/ARMCANZ (2000) document and remedial action suggested. One of the more interesting aspects of this statistical approach to setting a hazardous concentration was Australia and New Zealand's *implicit* support[1] for the use of a no observed effect concentration (NOEC) at a time when an OECD-sponsored workshop had recommended that the use of this statistic be discontinued (OECD 1998).

The NOEC has been a pivotal quantity in the whole theoretical development of SSDs and continues to be used in this context despite a plethora of papers highlighting its many and severe shortcomings (van der Hoeven 2004, Fox 1999). Indeed, Kooijman (1996) went so far as to say that "the NOEC should be banned". It is not my intention revisit the arguments for and against the use of NOECs here in any depth. What I do wish to draw attention to is, what appears to be, an increasing use of other measures such as the EC_{10}/IC_{10} either as a surrogate for, or as an alternative to the NOEC. Recent examples can be found in ecotox studies for Victoria's desalination project (Hydrobiology and CSIRO 2008) and on wastewater impacts from the Ranger Uranium mine (Hogan et al. 2008). In the remainder of this article I argue that this practice is flawed, both conceptually and operationally and suggest that NOECs be replaced by empirical estimates of model-based no effect concentrations (NECs). I am not the first to make this suggestion and indeed Koojiman et al. (1996), Jager et al. (2006), and others have argued the same point.

Most practising ecotoxicologists are well aware of the shortcomings of the NOEC. The most often cited concerns include: the NOEC is constrained to be one of the test concentrations; the procedure by which a NOEC is determined "rewards bad experiments"; statements of precision / uncertainty are not possible; NOECs cannot always be determined; and the size of the NOEC is a function of the choice of statistical test and level of significance. It is important to understand that the NOEC is a surrogate for the NEC and is routinely determined as the concentration in a series of dilution experiments for which the mean response is statistically indistinguishable from the mean response of a 'control' group. The statistical procedure most often used to assess the significance of differences between the control response and responses at all other concentrations is Dunnett's test (Dunnett 1955). Dunnett's test is a special case of a more general class of procedures referred to as multiple comparison techniques. Multiple comparison techniques are a companion tool for analysis of variance tests and should only be used *after* the null hypothesis of the equality of several means has been rejected by the ANOVA procedure. The rationale is that the multiple comparison tests will help identify the source of the significant ANOVA result. The analysis of variance technique was a stroke of statistical genius – it allowed the efficient testing of a hypothesis of the equality of a number of *means* by using a 'backdoor' approach based on an examination of components of *variance*. It is, nevertheless a fairly blunt instrument that can only conclude that a group of k means are either all the same (the null hypothesis) or at least two means are different (the alternative hypothesis). Rejection of the null hypothesis tells us nothing about *which* means are different.

Multiple comparison procedures such as Fisher's test, Tukey's test, Student Newman Keuls test, Hsu's test, and Dunnett's test are all based on pairwise comparisons of means to explore these differences. The test procedures differ primarily in terms of which error-rate is being controlled for (e.g. individual Type I error rate, overall 'experimentwise' error rate, etc.). Dunnett's test focuses specifically on the *k-1* comparisons of

*Author for correspondence, email: david.fox@unimelb.edu.au

treatment means with the designated 'control' group mean and does not concern itself with comparisons among pairs of non-control responses. Given that many concentration-response experiments employ a geometric progression of dilutions (e.g. a doubling of successive concentrations) then it is easy to see that the NOEC could be in error by up to the same constant of proportionality (e.g. by a factor of 2). Notwithstanding the other important and somewhat neglected ANOVA assumptions of independence, constant error variance, and, to a lesser extent, normally-distributed responses, why is this the preferred way of estimating the NEC? Why, when we have the opportunity of *modelling* the concentration-response data from which we can directly *estimate* the NEC, do we elect to use a less efficient and dubious multiple comparison test procedure with all its acknowledged faults? Perhaps it's because there's only one way of performing Dunnett's test whereas there is a multitude of concentration-response models, thereby imparting an assumed element of standardisation. I don't find this a compelling argument.

Multiple comparison procedures are wasteful of information, they are not predicated on any understanding of the system / experiment, and accordingly they represent a dumbing-down of ecotoxicology. I'm not alone in my harsh assessment. Nelder (1971, 1999) was more strident when he claimed that "multiple-comparison methods have no place at all in the interpretation of data" in his prescient article on statistical practice. Nelder attributes the widespread use of such "non-scientific statistics" to an obsession with *p*-values leading to "the cult of the single study and the proliferation of multiple-comparison tests". He makes a convincing argument for increased focus on *modelling*, claiming the basis for a "good" model is one that is: (i) *a priori* reasonable; (ii) parsimonious; and (iii) internally consistent. I believe that concentration-response modelling should focus on these three qualities rather than the routine application of unstructured statistical tests. This view is consistent with the calls for the use of biologically-based models such as those based on Dynamic Energy Budget theory (Kooijman 2006, OECD 2006).

This brings me to the issue at hand: are the problems with the NOEC and the associated statistical methodology overcome or ameliorated by using a different measure, such as the EC_{10} or IC_{10}? It may well be that the estimation of the EC_{10} is more reliable by virtue of the fact that we are attempting to estimate something less extreme than the NEC, but there is a fundamental schism that renders the derived SSD difficult to interpret. The schism is explained by the terms themselves: one relates to an effect; the other relates to *no effect*. While there is nothing to prevent us from taking a collection of EC_{10} values, fitting a *log-logistic* distribution say, and then using a low-order percentile from this fitted distribution as a threshold concentration, it does generate both a philosophical and operational dilemma. If one accepts the definition of a species sensitivity distribution as being the probability distribution of *some* measure of toxicity, then this procedure generates a SSD from which an HC_p could be determined. More generally, we can talk of the HC_p determined from the SSD based on EC_x data as being the concentration having an effect of no more than *x%* on at least (100-*p*)% of all species. However, except for *x=0*, I find this an awkward and convoluted concept. Indeed, what fraction of the population is protected by keeping environmental concentrations below an HC_p which has been determined from the SSD fitted to EC_x data and who decides on the value of *x*? The questions become even more difficult to answer if, as is often done, we introduce the notion of a $(1-\alpha)100\%$ confidence limit on the estimated HC_p. Based on an assumption of a log-normally distributed SSD, Van der Hoeven (2004) showed that when the variance between species is larger than the variance within a species, a relatively large portion of affected species, are affected severely.

As stated at the beginning of this article, it appears that there is an increasing tendency to use quantities such as the EC_{10} and IC_{10} as surrogates for the NOEC (which itself is a surrogate for the NEC!). In a recent analysis of the impact of waste water from the Ranger Uranium mine, Hogan et al. (2008) cited van der Hoeven (1997) to justify their use of low effect IC values on the basis that these were more robust than NOECs. In Part III of three related papers, van der Hoeven (1997) certainly recommended that "the NOEC should be gradually replaced by an EC_x estimation". However the advice was equivocal, with van der Hoeven stating earlier in the same paper that "if an NEC exists, ideally that is the parameter we want to estimate" and in Part I van der Hoeven et al. (1997) recommended that in addition to the EC_x, the parametric NEC (i.e. a model-derived estimate) be seriously considered.

In another recent example, Warne and Hobbs (2008) used EC_{10} data in their evaluation of the toxicity of the waste stream from the proposed Wonthaggi desalination plant in Victoria. In justifying this choice, Warne and Hobbs (2008) noted "we are using EC_{10} data whereas usually NOEC data are used which correspond to a EC_{10} to EC_{30}". While this may well be the case, to the lay person, this is a confusing statement as it seems to equate no effect with some effect. The source of confusion lies in the assumed equivalence of a NOEC and NEC (see accompanying article by Warne and van Dam). A no effect does not correspond to an effect (of any magnitude, except zero!). While it is entirely possible that a *confidence interval* for the NEC may include concentrations at which an effect is possible, this does not establish a correspondence between a NEC and an EC_x and certainly does not provide a reason for substituting one for the other. To do so, only results in a further obfuscation. A more detailed justification of the use of the EC_{10} was given in Appendix 1 of the report by Hydrobiology and CSIRO (2008). The Hydrobiology and CSIRO (2008) document claims that NOEC is a misleading term because it is defined as the highest concentration that causes an effect which is statistically indistinguishable from the control(s) and thus does not correspond to 'no effect'. I don't believe this is entirely accurate. The NOEC is the largest concentration at which the observed mean response is not statistically different from the mean response of the control group. Whether or not the control response represents an effect or no effect is not considered. It is not uncommon, for example to see *some* mortality at a control concentration representing a zero concentration of the test chemical. This

might be due to factors totally unrelated to the experiment, such as natural mortality or attrition. For the statistician, such outcomes represent 'noise' around a true response. Perhaps the confusion could be removed by talking about the Control Response (CR) rather than no effect. The outcome from Dunnett's test (if it must be used) is the largest concentration at which there is no statistical difference from the CR. Thus the NOEC could be replaced by the IFCR (indistinguishable from control response) or NDFCR (not different from control response) – but I suspect we don't need any more acronyms!

So, where does this leave us? For me, the answer is clear: invoke the principles (i) to (iii) above as advocated by Nelder (1999) and adopt a model-based approach to describe the fundamental concentration-response mechanism and the rest follows. The NEC and EC_x are respectively, a parameter estimate and a model prediction from one and the same model. Uncertainty in these values is handled with confidence and prediction intervals and more recently, Fox (in press) has described a Bayesian approach for setting credibility intervals using posterior and predictive distributions. The old method of using an unstructured, uninformed, and insensitive multiple comparison procedure is a bankrupt approach that deserves to be buried. Only then can we move forward and focus on more interesting modelling and estimation issues rather than trying to find ways to prop up the thoroughly flogged, dead NOEC horse.

ACKNOWLEDGEMENTS

I am grateful to Dr. Graeme Batley, Dr. Jenny Stauber (CSIRO Land and Water) and two anonymous referees for their helpful comments in preparing this article.

REFERENCES

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.

ANZECC. 1992. *Australian Water Quality Guidelines for Fresh and Marine Waters*. Australian and New Zealand Environment Conservation Council, Canberra, Australia.

ANZECC/ARMCANZ. 2000. *Australian and New Zealand Guidelines for Fresh and Marine Water Quality*. National Water Quality Management Strategy Paper No 4, Australian and New Zealand Environment and Conservation Council / Agriculture and Resource Management Council of Australia and New Zealand, Canberra, Australia.

Dunnett CW. 1955. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50**, 1096-1121.

Fox DR. 1978. *Mathematical Models of the Victorian Power Production System of Victoria*. Unpublished Masters Thesis, Monash University, Clayton Victoria, Australia.

Fox DR. 1999. *Setting Water Quality Guidelines – A Statistician's Perspective*. SETAC News, 17-18, May 1999.

Fox DR. (*in press*). An Bayesian approach for determining the no-effect concentrations and hazardous concentration in ecotoxicology. *Ecotoxicology and Environmental Safety*.

Hogan A, van Dam R, Houston M and Lee N. 2008. *Toxicity of Ranger Mine RP2 and Pit 3 Waters to Native Freshwater Species: 2007 Wet Season.* Supervising Scientist Report 197, Supervising Scientist, Darwin NT, Australia.

Hydrobiology and CSIRO. 2008. *Toxicity Assessment for the Victorian Desalination Plant*. Hydrobiology & CSIRO, Australia.

Jager T, Heugens EHW and Kooijman SALM. 2006. Making sense of ecotoxicological test results: towards application of process-based models. *Ecotoxicology* **15**, 305-314.

Kooijman SALM. 2006. An alternative for NOEC exists, but the standard model has to be abandoned first. *Oikos* **75**, 310-316.

Kooijman SAL, Hanstveit AO and Nyholm N. 1996. No-effect concentrations in algal growth inhibition tests. *Water Research* **30**, 1625-1632.

Nelder JA. 1971. Discussion on the papers by Wynn and Bloomfield, and O'Neill and Wetherill. *Journal of the Royal Statistical Society. Series B, (Methodological)* **33**, 244-246.

Nelder JA. 1999. Statistics for the Millennium: From statistics to statistical science. *The Statistician* **48**, 257-269.

OECD. 1998. *Report on the OECD Workshop on Statistical Analysis of Aquatic Toxicity Data. Series on Testing and Assessment, No. 10.* Environmental Health and Safety Publications, Series on testing and assessment. ENV/MC/CHEM(98)18.

OECD. 2006. *Current Approaches in the Statistical Analysis of Ecotoxicology Data: A Guidance to Application. Series on Testing and Assessment,* No. 54. Environmental Health and Safety Publications, Series on testing and assessment. ENV/JM/MONO(2006)18.

Shao Q. 2000. Estimation for hazardous concentrations based on NOEC toxicity data: an alternative approach. *Environmetrics* **11**, 583-595.

Tadikamalla PR. 1980. A look at the Burr and related distributions. *International Statistical Review* **48**, 337-344.

van der Hoeven N, Noppert F and Leopold A. 1997. How to measure no effect. Part I: Towards a new measure of chronic toxicity in ecotoxicology. Introduction and Workshop results. *Environmetrics* **8**, 241-248.

van der Hoeven N. 1997. How to measure no effect. Part III: Statistical aspects of NOEC, EC_x, and NEC estimates. *Environmetrics* **8**, 255-261.

van der Hoeven N. 2004. Current issues in statistics and models for ecotoxicological risk assessment. *Acta Biotheoretica* **52**, 201-217.

Warne MStJ. 1998. *Critical Review of Methods to Derive Water Quality Guidelines for Toxicants and a Proposal for a New Framework*. Supervising Scientist Report 135, Supervising Scientist, Canberra, ACT Australia.

Warne M and Hobbs D. 2008. "*The Victoria Desalination Project - Whole Effluent Toxicity Testing Results*". Presentation to Victorian Desalination Project Directions Hearing, 8 October 2008 Wonthaggi, Victoria. http://dsedocs.obsidian.com.au/planning/victorian-desalination-project/hearing-documents/023 michael warne- csiro overhead presentation.pdf

[1] Warne (1998) did in fact discuss the limitations of the NOEC and LOEC statistics but noted that no regulatory body had recommended their use be discontinued.