

***EcoStat, Inc.***  
P. O. Box 425  
Mebane, N. C. 27302

Ph/Fx: (919) 304-6029

[billwh@mindspring.com](mailto:billwh@mindspring.com)

May 17, 2007

Mr. Joe Karkoski  
Senior Water Resources Engineer  
Central Valley Water Quality Control Board (CVWQCB)  
11020 Sun Center Drive  
Rancho Cordova, CA 95670-6114

Subject: Review of *Methodology for Derivation of Pesticide Water Quality Criteria for the Protection of Aquatic Life in the Sacramento and San Joaquin River Basins. Phase II: Methodology Development and Derivation of Chlorpyrifos Criteria*

Below, please find my review of the subject report. The document contains numerous statistical flaws that are not consistent with good statistical practice. Unfortunately, these flaws negate the credibility of the document. I have confined my comments to the statistical aspects of the subject report in Chapters 2 and 3. The comments are numbered sequentially, below.

## **Chapter 2**

1. p. 1-2, Table 1.1: The authors reviewed existing methodologies and finding them lacking, derived their own. However, Table 1.1 is incomplete and does not represent the current approaches for evaluating pesticide toxicity data. The table is focused on a selected few methods used by regulatory agencies rather than a compilation of the newest and most innovative methods. For example, the methods contained in the following examples do not have the statistical flaws contained in the subject document. There are many interesting and innovative methods for establishing criteria that are not referenced by the authors. Some recommended approaches are included in the following documents:

- a. The USEPA co-sponsored a Pellston conference entitled *Application of Uncertainty Analysis to the Ecological Risk of Pesticides* that produced many papers and approaches for establishing criteria. (In preparation for publication by SETAC Press, Pensacola, FL)
- b. *Species Sensitivity Distributions in Ecotoxicology*. 2002. Leo Posthuma, Glenn Suter, Theo Trass. eds. Lewis Publishers, New York.
- c. Warren-Hicks, W. J., D. Moore. eds. 1998. *Uncertainty Analysis in Ecological Risk Assessment: Pellston '95*. SETAC Press, Pensacola, Florida.

- d. J. B. Parkhurst, Warren-Hicks, W. J., R. Cardwell, J. Volosin, T. Etchison, J. Butcher, S. Covington. 1995. Risk Managing Methods: Aquatic and Ecological Risk Assessment Aids Decision-Making. Water Environment & Technology, November 1995.

2. p. 2-3, Hypothesis tests vs regression analysis: In this discussion, the authors confuse concepts in statistics with concepts in laboratory testing. Methods used to analyze toxicity data should not be confused with quality assurance (QA) criteria that evaluate acceptable test results. Once laboratory data are deemed acceptable based on indicators of good laboratory practice, the data should be considered appropriate for statistical analysis. The results of a statistical analysis should not be used to judge the data invalid. For example, a high minimum significant difference (MSD) does not necessarily reflect a “bad” test, but may reflect the fact that the toxicant is relatively weak resulting in a large between-concentration variance. A properly constructed MSD reflects the data; it does not invalidate the data as indicated by the authors.

There are a very few cases in which the output from a statistical analysis can be used as a QA tool for toxicological data. Statistical methods can be used as a tool to evaluate the presence or absence of the underlying conceptual model. For example, toxicological theory supports the concept that toxicity should increase with concentration. Statistical tests of monotonicity may be appropriate because they reflect the underlying model. [I note that monotonic dose-response curve issues are not considered here.] If the objective of the toxicity test is to identify say, an IC25, and the data are not sufficient for this purpose, then the test information may not be useful. Even in this case, however, providing the data meet basic laboratory QA criteria, the data themselves should not be labeled as invalid.

A large within- or between-concentration variance does not invalidate the data, nor does it invalidate the subsequent analysis of the data. QA decisions should not be solely based on the size of the variance estimator. It is important in this report that issues of quality assurance and statistics are not melded as they appear to be currently.

Note that model-based endpoints and hypothesis-based endpoints have many of the same interpretation issues. The authors state that: “Although regression methods are preferred, there is little agreement among scientists as to what level of statistical effect may be considered a no biological effect ...” I note that this is the same issue surrounding the NOEC-LOEC-MATC issue that the authors discuss on the next page. The bottom line is that establishment of a statistical endpoint for setting criteria is arbitrary, whether the method is model-based or hypothesis based.

I agree that model-based endpoints are preferred. It would be helpful if the authors could provide literature citations on this issue at this point in the document.

It is also recommended that the authors point out in this section that model-based endpoints and hypothesis-based endpoints cannot be combined in the same analysis. These endpoints reflect a completely different underlying conceptual model. See more comments below on this issue.

The concept of a reliability score is inappropriate. See comments below.

3. p. 2-12, Ecotoxicity data evaluation: The rating techniques described in this section are inappropriate, have no viable interpretation, and do not represent acceptable quality assurance practice and procedure. Toxicity information, from multiple sources, should be judged based on the adherence to standard practice or guidance concerning the conduction of the test. Once the data pass these basic laboratory-based QA criteria, the data are then sufficient for statistical analysis. The weighting scheme described in this section is arbitrary, and the interpretation on p. 2-14 (based on percentiles) has no scientific basis. For example, why are relevance scores greater than the 90<sup>th</sup> percentile rated “relevant”? Why not use the 85<sup>th</sup> or 95<sup>th</sup> percentile? This approach, based on questionable data management and poor statistical practice, is scientifically weak.

Note the following sentence, “The 75<sup>th</sup> percentile of scores is suggested for the reliability rating because, in the case of chlorpyrifos data set, higher percentiles were too restrictive, resulting in rejection of too much data ...” This sentence illustrates the issues I have with this section of the document. Good quality assurance criteria and procedures are not established based on the amount of data remaining for analysis. Data should not be discarded based on an arbitrarily established statistical endpoint. This approach to data reduction is inappropriate. Data should only be discarded if they are “wrong.” An investigator may decide that data are not appropriate for the model under analysis, but the data should not be labeled “unreliable,” since the data may be appropriate for other analyses.

Again, the authors of this report are confusing good statistical practice with issues of basic data quality. This method results in a loss of information that may not be required from the perspective of identifying viable information. This section and associated procedures should be discarded.

4. p. 2-14, Sample size: Where did a sample size of 5 tests come from? A correct evaluation of sample size should consider the following: (1) the model under evaluation, (2) the variance of the data (or model terms), and (3) the pre-defined requirements for accuracy, precision, and acceptable error. In this document, the sample sizes are established out of convenience. Approached scientifically, a “new” method would explore issues associated with precision, accuracy, and error requirements in a formal manner before blessing a sample size.

In any case, the objective of this document is a methodology for water quality criterion development. It is inadvisable to develop regulatory criteria from small data sets (n=5, for example). Therefore, based on good scientific judgment and practice, if data are lacking then the methodology should simply require additional data.

I strongly suggest that the authors conduct a formal sample size analysis before defining the number of toxicity tests required for criterion development. The authors should pay careful attention to issues associated with within- and between-laboratory variance associated with various toxicity test endpoints. There is a rather large literature on this subject, and again, a SETAC Pellston conference dealing with this issue.

5. Section 2-2.7, Data reduction: I understand the issue of over-weighting the relative toxicity associated with a specific species. In addition, I have no problem with the general guidance provided in this section. I suggest, however, that the authors add narrative that addresses issues associated with combining the various toxicity endpoints (e.g., IC25, NOEC, LC50) within a single SSD. I

strongly believe that a single SSD can only be comprised of data representing a single toxicity endpoint. This issue should be clarified and discussed in this section.

6. p. 2-19: The authors state that, “The aim of both SSD and AF methods is to extrapolate from available toxicity data for a limited number of species to toxicity values that will be protective of all species in an ecosystem.” Is this really the goal? If so, why not set the criteria to zero? Many regulatory agencies (e.g., the USEPA) recognize that complete protection is not possible, or even useful.

7. Section 2-3.1.1, Appropriate distribution: This section does not reflect good statistical practice and modeling. The stated reasons for selecting a distribution are inappropriate. Investigators should not choose a distribution because “... how many samples are required, on which distributions are easier to work with, or which ones better quell the criticism that SSDs are not valid ...” I note that this sentence represents a misunderstanding of good statistical practice. Also, I note that the USEPA’s use of a triangular distribution, which has no interpretation within a toxicological paradigm, is also inappropriate. I encourage the authors to provide guidance for distribution selection that is consistent with good statistical practice.

Selection of an appropriate distribution should be based on the underlying conceptual model. For example, for acute data (life/death) binomial distributions of survival are appropriate. In this case, a generalized linear model using a log-logistic link function linking the concentration response data to the probability of survival is consistent with the underlying conceptual model and the data collected to parameterize the model. Extensions of this model that lead to the mathematics underlying SSDs comprised of acute (binary) response metrics can be found in the references provided at the top of this report. Distributions for continuous data or counts should be appropriately chosen. I refer the authors to a long series of papers written by A. J. Bailer and J. T. Oris for further examples of proper statistical approaches for developing models and distributions with toxicological data.

In no case should distributions be selected because they are convenient, easy to use, or match the available sample size (like the triangular). The distribution must represent the toxicological and biological process of interest and be mathematically tractable within that process.

Selection of a distribution based on the conceptual model is a major reason why it becomes difficult to merge acute and chronic data into a single SSD. The endpoints are mathematically derived from different conceptual and mathematical models, making their cross-interpretation at best difficult, if not impossible. For example, it is difficult to compare an IC<sub>25</sub> derived from an underlying binomial process with, say, a NOEC derived from an underlying log-normal process. Not only are the statistical methods used to generate these endpoints not comparable, the interpretation of toxicity inferred by each is incompatible.

As the field of environmental toxicology has matured, the acceptance of the link between distributions and biological interpretation has evolved. In the human health sciences this is equivalent to the interpretation of a gamma function for survival-time studies. Therefore, I see no need to involve another class of distributions (i.e., the Burr III family) without a thorough understanding of the link between the mathematics and biological and toxicological processes. I am

not aware of published studies on how the Weibull and Pareto distributions are interpreted within the context of toxicological information. In fact, these distributions would be inappropriate for binary response metrics.

I strongly suggest that this entire section be rewritten. The references at the top of my review have a great deal of information on how to correctly select and defend a distribution.

8. p. 2-28: Choice of distribution should not be made based on statistical goodness of fit tests. The distribution must be interpretable within the process under evaluation. If the distribution does not fit the data, then the investigator may need to rethink his/her original hypothesis. But, the fit statistic should not be used to establish the underlying model (as is the case in this section). Please see my comments above. Also, I refer the authors to the above referenced documents on uncertainty analysis, which contain discussions of the misuse of curve fitting methods in decision-making. This section of the report should be rewritten to reflect the current literature in the statistical analysis of toxicological data within a risk paradigm.

9. p. 2-37, Percentile cutoff: The authors provide a reasonable discussion of this issue. I strongly urge the authors consider the statistical methods used in the Water Environmental Research Foundation (WERF) risk methods and software (cited above). The methods for distribution choice provided in the WERF documentation are consistent with good statistical practice. Furthermore, the WERF method presents approaches for setting SSD-based criteria using uncertainty estimates. The method provides derivations of all statistical functions at fixed percentiles of the SSD, thus eliminating the need for safety factors.

10. p. 2-39, Aggregation of taxa and outliers: Again, the distribution of choice must reflect the biological and toxicological process under evaluation. If modalities are evident because of differing biological or toxicological processes, then I agree that data should be separated. They are separated because the underlying process is not consistent with the conceptual model reflected by the choice of distribution.

However, never eliminate data because they do not fit the model well. Valid data should be used, even if they do not represent the choice of distribution. Uncertainty in the data is informative and should not be eliminated. The author's statement that "... it is reasonable to exclude outliers ..." represents poor statistical practice.

11. p. 2-39, Comparison of Methods: From a theoretical perspective, only the methods of Aldenberg & Jaworska (2000) overcome some of the many issues raised above. This method can be adapted for binary, continuous, and cardinal data. The other methods are lacking in either mathematical rigor, flexibility, or interpretation. The authors will find that the Aldenberg & Jaworska approach is similar to the WERF methodology.

12. p. 2-45, AF methodology: In this age of modern computing and internet communication, there should be no reason to use safety factors (a.k.a. assessment factors) - for any reason, under any conditions. The literature is replete with mathematical methods for calculating uncertainty in SSDs and concentration-response models. The use of safety factors is simply to assure policy makers that their resulting criteria are protective. However, as the narrative correctly notes (but for some reason

then ignores), there is no safety in the use of safety factors. Their use is not a reasonable approach for predicting the future protectiveness of the criterion. Effectively, the use of safety factors negates the influence of science and mathematics in policy decisions.

Criteria should not be developed from small data sets. Therefore, the use of these factors is not justified based on first principles.

13. p. 2-52, ACRs: If the objective is to develop a water quality criterion with regulatory implications, then chronic data should be generated. Substituting the use of an ACR in place of generating chronic data is not appropriate for establishing standards with regulatory implications. The literature contains numerous studies (some referenced in the document) concerning the large range of ACR values for a single chemical/test species/endpoint combination. Therefore, ACRs should not be used, and chronic data should be generated when needed. Appropriate methods for selecting the number of chronic tests is addressed in my comments above.

14. p. 2-55, Averaging period: The statistical issues underlying an appropriate averaging period, and the number of water quality samples required during the averaging period, include the following: (1) temporal variability in pesticide concentration, (2) occurrence of temporal correlation (i.e., autocorrelation or seasonal patterns), and the statistic of interest (mean, upper percentile, etc.). The USEPA has regulatory guidance that was developed without the explicit analysis of these issues within the context of criterion setting. Approached scientifically a “new” method should address these issues when attempting to establish a time-period for water quality sampling. Without addressing these issues, the authors risk both false positive and false negative results within their regulatory framework. I suggest that the authors formally address the averaging time and associated sampling issues prior to endorsing an approach.

15. Section 2-3.5.2, Mixtures: It is proposed that the authors consider the WERF methods, and papers presented at the SETAC Pellston conference on Uncertainty Analysis of Pesticides, prior to selecting an approach to dealing with mixtures. These references present formal mathematical approaches for combining data across species within a single chemical, and across chemicals within a single species. The methods presented in this section lack a proposed approach for dealing with uncertainty. Also, these methods focus on a single effects endpoint rather than the entire concentration-response curve. A “new” method should provide insights into more advanced methods for combining data.

### **Chapter 3**

Please note that each of the issues addressed above is reflected in the methodology described in Chapter 3. The authors emphasize the toxicological and biological issues underlying the proposed methodology over the statistical issues. However, the statistical issues underlying the method drive the selection of pesticide criteria.

Because Chapter 3 has the same basic structure as Chapter 2, I simply reiterate my major concerns for this methodology:

1. The Burr III family is inappropriate for binary response data. Furthermore, the Weibull and Pareto distributions have no interpretation within a toxicological context. This is comparable to the dangers associated with Monte Carlo software (a highly misused tool) where the investigator is free to select interesting distributions without any knowledge of the underlying mathematics or associated links to the biological process under evaluation.

It is recommended that the authors provide sound mathematical and statistical arguments for their selection of the Burr III distributions. In particular, arguments linking the underlying biological process with the mathematics of the selected distribution should be provided. Currently, the document is lacking this defense. It is not enough to simply find an equation that generates a sigmoidal curve.

2. The relevance scoring system should be discarded and replaced with sound quality assurance criteria and practice.

3. Excluding data simply because of fitting issues is poor statistical practice. Outlier tests (Sokal & Rolf) should not be used as the basis for discarding data that have passed rigorous laboratory derived quality assurance criteria.

4. A formal analysis of the number of tests required for criterion setting should be developed and presented.

5. Assessment factors should be replaced with formal uncertainty analyses.

6. It is recommended that the authors re-consider their approach to working with mixtures. The references provided above are useful in this regard.

7. A formal analysis of the statistical issues underlying the selection of an averaging period and sample size associated with water quality sampling should be developed and presented prior to publishing this document.

Sincerely,

*submitted by email*

William Warren-Hicks, Ph.D.  
CEO