Peer Review Comments - Lisa Nowell, Ph.D., U.S. Geological Survey

Peer Review Questions

Methodology for Derivation of Pesticide Sediment Quality Criteria for the Protection of Aquatic Life. Phase II. Method Development and Derivation of Bifenthrin Sediment Quality Criteria

General comments:

The methodology clearly represents a huge and worthwhile effort and it incorporates state-of-the-science methods for estimating bioavailability. The methodology pushes the envelope in terms of sediment toxicity testing. I am eager to know whether any pesticides have sufficient data available to compute criteria using the SSD method proposed in this report; if not, the fallback method (which uses generic empirical assessment factors to compute acute criteria and acute-chronic ratios to calculate chronic criteria) would be used. Also, the use of Koc values (which have high variability) to convert between sediment organic carbon and freely dissolved porewater concentrations, or vice versa, adds substantial uncertainty to criteria values. As a result, uncertainty in calculated BSQC levels is likely to be high. In this case, I would encourage the use of BSQC as screening levels, but I question whether they are suitable for use in compliance monitoring, as was suggested by language in the report.

I suggest that a tiered structure may be appropriate, in which a clear distinction is made between criteria with lower uncertainty (e.g., determined using the SSD method using substantial toxicity test data representing appropriate taxa, endpoints, and test durations, and with minimal uncertainty introduced by use of any Koc, default ACRs or AFs), which may be suitable for compliance monitoring, vs. criteria with high uncertainty that were generated using assessment factors, ACRs, and/or geometric mean Koc values, which would be suitable for screening-level assessment. One important and useful component of a criterion (or any benchmark) is a statement of what it means when the criterion/benchmark is exceeded, and I recommend that you address this explicitly in your final criteria statement. This would greatly facilitate the application of the criteria by environmental regulators and researchers.

1. Is the way the method addresses bioavailability in accordance with the current state of research on this topic?
   Yes—the method pushes the envelope on bioavailability testing because most monitoring studies typically measure pesticides in whole sediment and report in units of sediment organic carbon. Because relatively few studies have measured pesticide toxicity/concentrations using the Tenax or micro-extraction methods, the method is probably ahead of the body of scientific literature at this point in time.

2. Are all of the ways of accounting for bioavailability included in the method (and listed below) scientifically valid? Are there additional technically valid ways to account for bioavailability that could be used?

    a. OC-normalized sediment concentrations
    b. DOC-normalized porewater concentrations
    c. Directly measured freely dissolve porewater concentrations (via SPME or Tenax)

    These approaches are valid ways of estimating (not measuring) bioavailability.

3. Will environmental regulators and researchers be able to use existing toxicity and monitoring data included in the method to check compliance or does the method require that new techniques be used to generate new data?
    Doubtful. (1) Because of the large variability in Koc values for hydrophobic contaminants, the uncertainty introduced by converting from pore water concentrations to sediment organic carbon, or vice versa, will be very high. To avoid this, the method used to measure concentrations in environmental samples would have to match exactly the method used to measure concentrations in toxicity tests used to determine the criteria for a given compound. More studies are needed that measure pesticides concurrently using microextraction, Tenax, and organic carbon-normalized sediment methods. (2) There are standard methods available for relatively few freshwater sediment taxa (largely midge and amphipods). The requirement for 5 families means that effort needs to be put into (a) development of standard methods for benthic organisms other than midge and amphipods, and (b) sediment toxicity testing of pesticides with appropriate nos. of taxa, in which pesticide concentrations are measured using multiple techniques (microextraction, Tenax, and organic carbon-normalized sediment concentrations). (3) Given the lack of sediment data for pesticides (and sediment standard methods), I wonder whether there are sufficient data available for any pesticides to develop a SSD using the protocol in this report? If there are example pesticide(s) with toxicity data for the 5 required taxonomic groups, then I think the method needs to be tested on such pesticide(s), and criteria values determined using both the SSD approach and AF approach (i.e., compare SSD-based criteria with criteria that would have resulted had only amphipod and midge data been available—as is typical for many pesticides, including bifenthrin). (3) The lack of chronic sediment toxicity data (and standard test methods for chronic tests) also means that ACRs will be needed to assess chronic toxicity. Standardized test methods are needed for additional test species, sublethal endpoints, and chronic test durations. (4) Use of default assessment factors and ACRs will result in highly uncertain criteria values, so criteria will likely be appropriate only for screening-level assessment. It seems premature to try and develop “criteria” with regulatory significance at this state of the science. I think the method you propose makes sense from a theoretical point of view, but there aren’t data enough to validate it. I suggest a tiered structure of guidelines/criteria, based on degree uncertainty (from curve-fitting, use of AF, use of ACR, variability in/use of Koc values to predict concentrations in porewater and/or sediment organic carbon, etc.)

4. Is it clear how to evaluate studies by reading section 2.3 and appendix A (rating guides) and looking at tables 7-13?
   Yes.

5. Do the categories and point values assigned in tables 8-12 reflect the importance of the parameters to performing valid sediment toxicity testing?
   It is not clear that they do. I would like to see you provide some justification for the relative points assigned. Some important design elements, such as control response, have few points assigned (poor control survival would result in loss of only 7.5 points in the Relevance score (Table 8) and loss of 6 points in the Acceptability score (Table 10), whereas detailed test conditions that are sometimes not reported in journal studies may result in substantial point losses in the Documentation score (Table 9) (e.g., 12 points for characteristics of overlying water). I am not saying these details are trivial, but does their omission necessarily preclude using the study in setting criteria? Why are they more important than poor control survival?

6. Is it clear how to prioritize and organize data by reading sections 2.4 and 2.5? Do the data prioritization and exclusion in the bifenthrin criteria derivation seem reasonable (section 8.7)? This step plays a large role in determining which data are used to derive the criteria, and thus the magnitude of the criteria.
   It is clear how to prioritize data, except as noted in my comments to the text. When you say that Tenax/microextraction data are preferred, does this imply that you would select them over tests that measured toxicity in units of sediment organic carbon, or would you combine these studies by converting the sediment-oc results to porewater concentrations using Koc values? I am not comfortable with the data exclusion rules, which may be straight-forward in terms of execution but may have unintended consequences. I am wary of picking the most sensitive endpoint for each taxon and mixing multiple endpoints/species in a SSD. Also, the selection of a nonstandard endpoint over multiple standard endpoints strikes me as problematic—this happened for bifenthrin, where the single instantaneous growth rate was selected over several studies with the standard growth and survival endpoints. It raises the possibility that a single study with a non-standard endpoint could form the basis of the criterion value (although this did not occur for bifenthrin).

7. Is it clear what information should be input in the toxicity data summary Table 14?
   Yes, although the order of information in Table 14 could be revised to make it easier to fill out Tables 8-11.

8. Are instructions in sections 3.4-3.7, describing how criteria are derived, clear and easy to follow?
   I did not completely follow description of calculation of Assessment Factors in 3.5.2 (step b). These sections contain a somewhat uneven mix of (important) background/supporting information and instructions. Section 3.4 gives clear

instructions for following the SSD procedure. Then section 3.5 gives background on the Assessment Factor approach and explains how the authors determined the default AFs, which vary as a function of the number of ecotoxicity requirements that have been met. The switch from instructions in Section 3.4 to background/methods development in 3.5 may be confusing to users trying to apply the method to develop criteria—especially because section 3.5.2 starts by saying. "The procedure used to calculate the acute AFs presented in the UCDSM is outlined in this section." This may suggest to the user that the user needs to calculate the acute AFs him/herself, whereas (if I understood correctly), all the user has to do is select appropriate AFs from Table 16. The instructions part of section 3.5 doesn't come until section 3.5.3. (Note that there is no comparable background on SSD curve-fitting, etc., given in section 3.4, and I think a summary of this would be appropriate to include in the report.) Section 3.6 is mostly instruction, with some background included. In general—it would be more user-friendly if you could separate important background from step-by-step instructions more clearly/consistently. For ex., perhaps you could put instructions in tables within each section or subsection.

Section 3.7 is confusing. Why are you talking only about chronic criteria for herbicides, and not acute criteria? You never describe the kinds of tests likely to be available for algae/macrophytes—are you talking about water tests? I do not understand why acute criteria for herbicides would be based largely on invertebrate test data, while chronic (only) criteria for herbicides would require tests with algae and macrophytes.

9. Does it make sense to derive two criteria for a given pesticide, one with a 10-d averaging period and one with a 28-d averaging period (section 3.8.2)? Should only one criterion be derived? Please comment on the thoroughness, validity, and completeness of the review and discussion in section 3.8.2. Are there are any other considerations that should be included for determining criteria averaging periods? The averaging periods make sense in that they are matched with duration of 10-d vs 28-d toxicity tests. But I don't see a good relation to the literature reviewed on temporal variability of pesticides in sediment. That literature is modest, and it is not possible to separate spatial variability from temporal variability due to pesticide application events, precipitation or irrigation events, or flow changes (scouring) within the stream. The review did not appear to help you select averaging periods because ultimately, you were constrained to select averaging periods that corresponded to the duration of standardized tests.

10. Is the assumption of concentration addition reasonable for mixtures of pesticides in the same class (section 4.2)?

    I think it makes sense to apply concentration addition to toxicity values (e.g., EC50 or EC20) for mixtures of chemicals with similar modes of action and dose-response relationships. I do not think it makes sense to apply it to criteria values, which may represent HA5 values from SSDs (where different species may have different endpoints, MOA, and dose/response relationships), may be calculated using different methods for different compounds (e.g., one compound's criterion may be the HA5 from SSD whereas another compound's criterion may be based on the lowest SMAV), or may represent SMAVs for different species and/or endpoints, and may incorporate different Assessment Factors.

11. Do you know of QSARs that could be used to estimate toxicity to other species, including threatened/endangered species?
    You probably know these, but:

Lessigiarska I, Wortha AP, Sokull-Klüttgen B, Jeram S, Dearden JC, Netzeva TI, Cronin MT, 2004. Qsar investigation of a large data set for fish, algae and Daphnia toxicity. SAR QSAR Environ Res. 2004 Oct-Dec;15(5-6):413-31.

Sala S., Migliorati S., Monti G.S. and Vighi M., SSD-based rating system for the classification of pesticide risk on biodiversity, Ecotoxicology 21, 2012, 1050–1062.

U.S. Environmental Protection Agency. Ecological structure activity relationships; http://www.epa.gov/oppt/newchems/tools/21ecosar.htm.

von der Ohe, P. C.; Kühne, R.; Ebert, R. U.; Altenburger, R.; Liess, M.; Schüürmann, G. Structural alerts—A new classification model to discriminate excess toxicity from narcotic effect levels of organic compounds in the acute daphnid assay. Chem. Res. Toxicol. 2005, 18 (3), 536–555.

Walter, H., 2002, Dissertation, cited in Walter, H., Consolaro, F., Gramatica, P., Scholze, M., Altenburger, R., 2002. Mixture toxicity of priority pollutants at no observed effect concentrations (NOECs). Ecotoxicology 11, 299–310.

12. Are the bifenthrin criteria generated in section 8 protective of aquatic life, more specifically, are they neither unreasonably overprotective nor underprotective?
    I don't see how we can evaluate whether the bifenthrin criteria are protective, overprotective or underprotective. Because AF and ACR methods were used, they may well be overprotective—but this remains to be determined. We don't have the data with which to assess this—we would need toxicity test data with multiple test species (including the 5 required taxonomic groups) in multiple sediments. The empirical AFs are based on water toxicity tests with 5 required taxonomic groups and represent the probability that an untested species may have a lower SMAV than the most sensitive species tested, as a function of the no. of species tested (*n* =1-5); I understand why it is necessary to use water data to compute these AFs. However, sediment toxicity data are often available for only 2 taxonomic groups (amphipods and midge), of which one (amphipods) is known to be highly sensitive to pyrethroids and many other insecticides. Therefore, the amphipod SMAV will often form the basis of the criterion using the AF approach. It makes sense to test whether the water AFs accurately reflect the probability of an untested species having a lower SMAV than amphipods in sediment (for *n* = 2 tested species). There may not be sufficient data to do this, but I think such a test is important to try before we can state how protective the BSQC are. Also, the observation of resistance to pyrethroids in local field populations of *Hyalella azteca* (Weston et al., 2013, PNAC, v. 110, p 19532) raises an interesting question—Is development of resistance in local

populations of Hyalella azteca that are exposed to pesticides considered to be an adverse effect of these pesticides on local communities? How does one factor this into use of benchmarks (such as BSQC) to assess/predict potential effects on aquatic communities?