External Peer Review of the Report on Relationships Among Exceedences of Chemical Criteria or Guidelines, the Results of Ambient Toxicity Tests, and Community Metrics in Aquatic Systems

Submitted to:

Michael Griffith
National Center for Environmental Assessment
U.S. Environmental Protection Agency
26 W. Martin Luther King Drive
Cincinnati, OH 45268

Submitted by:

Eastern Research Group, Inc.
110 Hartwell Avenue
Lexington, Massachusetts 02421-3136
February 20, 2006

QUALITY NARRATIVE STATEMENT

ERG selected reviewers according to selection criteria provided by EPA. EPA confirmed that the scientific credentials of the reviewers proposed by ERG fulfilled EPA's selection criteria. Reviewers conducted the review according to a charge prepared by EPA and instructions prepared by ERG. ERG checked the reviewers' written comments to ensure that each reviewer had provided a substantial response to each charge question (or that the reviewer had indicated that any question[s] not responded to was outside the reviewer's area of expertise). Since this is an independent external review, ERG did not edit the reviewers' comments in any way, but rather transmitted them unaltered to EPA.

TABLE OF CONTENTS

Section I: Individual Peer Review Comments

Section II: Comment Disposition Report

Appendix A

SECTION I

INDIVIDUAL PEER REVIEW COMMENTS

REVIEW COMMENTS OF REVIEWER 1

Jerome Diamond, Ph.D.
Director, Tetra Tech, Inc.
400 Red Brook Boulevard, Suite 200
Owings Mills, MD 21117
410-356-8993
Email: jerry.diamond@tetratech.com

Review of Draft Report, Relationships Among Exceedences of Chemical Criteria or Guidelines, the Results of Ambient Toxicity Tests, and Community Metrics in Aquatic Ecosystems

Introduction

1. *Does the introductory chapter make a coherent statement about the nature, purpose and limitations of this document, and of the research it describes?*

   Presents the nature and purpose fairly well but not the limitations. There is no material presented that would tell the reader how generalized the approach or results of this study are. Also, no real limitations presented in terms of the type of data available and data characteristics.

2. *For those areas within your expertise, is the information accurate, clear and concise?*

   This chapter and others need technical editing in many places. Some text is awkwardly constructed and difficult to follow (e.g., lines 9-12, p. 3). Specific technical suggestions are:

   (a) The EPA tox test references (p. 2) are old. If these are what were used in the study (as opposed to the higher QA/QC stipulated in more recent methods), then I don’t think the authors can claim use of current methods

   (b) Sediment toxicity effects are usually predetermined by comparing test sites to reference sites (see ASTM standards on this) not controls (line 17, p. 2). Controls are used for QA/QC purposes.

   (c) Metrics are one tool used to analyze bioassessment data - not the only tool as one would think reading the last paragraph p. 2. Multivariate tools are also used extensively (e.g., RIVPACS).

   (d) Lines 3-5, p. 3 are vague. It is not clear from this how metrics are to be used.

   (e) I disagree that chemical criteria are organism-based. They are calculated based on toxicity data but the criteria are meant to represent populations not organisms - hence the difficulty with using criteria for endangered species protection (which is organism-based).

   (f) References cited in lines 3-5, p. 4 are old. Should reference the Pellston SETAC book on this topic (Groethe et al., 1996) at least. Others are included in a special ET&C issue on WET (e.g., Diamond et al., 2000).

   (g) The use or lack of use of <u>many</u> community metrics is not necessarily an issue. The metrics need to be calibrated for the site and it is important that redundant metrics are removed.

3. *For those areas outside your expertise, is the information clear, concise and easy to follow?*

   All information was within my area of expertise.

4. *In terms of completeness, organization and level of detail, does the information seem to provide an appropriate introduction to the topics covered, for the purposes of this document*

   Level of detail is appropriate. Seems complete.

Southern Rocky Mountain Ecoregion R-EMAP Study

5. *Is this study presented clearly, and would it be easy to understand for a reader unfamiliar with the three study methods?*

   I realize the chapters were prepared as somewhat stand alone chapters, but if this is to be a report, there is no need to repeat the introduction given in Chapter 1. Pages 8 and 9 should be deleted. Start with first real paragraph p. 10.

   The study is presented fairly clearly and much of it would be understood by environmental scientists but probably not by a non-biologist. There are several things that I think should be clarified:

   (a) Figure 1 - include a scale so someone understands what size area was sampled.
   (b) Were there any other potential sources of stress for the seven “downstream” sites besides mining?
   (c) How did flows compare between 1994 and 1995? Demonstrate to the reader that combining the data from both years in one analysis is scientifically valid.
   (d) What was the variability between visits to a given site (line 3, p. 12)?
   (e) Should include table with method used and detection limits achieved for each metal analyzed.
   (f) What QC tests were conducted to support the toxicity tests? Reference toxicant tests?
   (g) Should include acceptability criteria used for each type of toxicity test.
   (h) The *Hyalella* test appears to be a non-standard test. Normally this is a 10-d test, not a 7-d test. Why was a 7-d test used? Growth is difficult enough to measure in a 10-d test, I find it difficult to believe that 90% growth of controls could be significantly different. The “preliminary comparisons” (line 21, p. 13) need more explanation and some back-up because much of the analyses depended on the thresholds determined from this comparison.

(i) Were the sediment samples for chemistry and toxicity taken from the riffle areas at which macroinvertebrates were collected and used in analyses?

(j) Tables 1 and 2 appear to be a laundry list of metrics. Given the preponderance of metals in this region as a main stressor, it would seem more fruitful to examine those metrics that have been shown to be responsive to metal stress. I'm concerned that when so many metrics are examined (there's more than 50 represented here), just by chance, you might expect to see significant differences for a few of them. At the very least, I would strongly recommend looking at the correlations among metrics and delete those that are redundant.

(k) Table 1 - I don't understand how the "D" column F values were derived when you have four different metals potentially. In general, I found it difficult to interpret the meaning of this table. As noted previously, if there were fewer metrics examined, and they were uncorrelated with each other, that would help the readability of this table.

(l) How come the authors didn't look at additivity of metals as one factor, rather than looking at each one individually in correspondence analyses? The latter approach inappropriately lumps together sites where only one metal exceeds a threshold with sites that exceeded multiple thresholds. There is also no accounting for how much greater the concentration at a site is compared to its threshold. This too adds uncertainty to the analysis.

(m) Disagreement assessment of metrics (lines 3-12, p. 20) is not clear for metrics that mean "worse" conditions as they increase in value.

6. *Are the findings and the limitations of the analyses correctly stated?*

I disagree with the interpretation of the findings in several places. First, it is unclear how much extremely contaminated sites are driving any relationships or regressions observed. A cumulative probability plot or a box plot of metal concentrations observed at all sites in the dataset should be done for the bioassay data.

Second, it is clear that most of the sites have low metal concentrations and no toxicity. The y analysis appears to be biased by the agreement in one cell of the matrix in each case. A true accurate assessment would remove those sites for which there are no apparent stressors present and look at the remaining relationships in which at least one measurement endpoint is exceeded. If this were done, the relationships between endpoints are pretty bleak. I would note that this problem is not unique and has been observed in recent studies including WET-bioassessment comparisons (e.g., Diamond et al. 2000).

Third, as the authors acknowledge, there are some large disconnects between the types of endpoints being compared; e.g., sediment thresholds based on 28d *Hyalella* tests versus results of 7-d *Hyalella* sediment bioassays. The fact that acute bioassay results were related at all to benthic metrics suggests to me that there must have been at least a few sites that were severely contaminated with metals (not unlikely in this region), and that results from these sites were driving the statistical relationships observed.

Fourth, I concur with the approach taken in the results and discussion, examining the extent to which chemistry and bioassay measures, in relation to their thresholds, agree with the bioassessment metric data. But the results of this approach only partially supports the authors' conclusions in my opinion. The authors point out the large discrepancy in assessment results using bioassays and chemical thresholds and they also point out the many discrepancies observed between assessments at sites based on metrics and other measures. The point is that on a site-by-site basis, results of different measures are often conflicting (i.e., disagree), negating the conclusion that "organism-level effects are predictive of effects at the community level" (line 2-3, p. 36 and lines 2-6, p. iii, Abstract). The report needs to clearly state that the relationships observed are statistical, based on the data in general, and NOT predictive of relationships at any given site.

Fifth, the conclusions presented for the piecemeal regression analysis appears unfounded to me based on the plots in Figures 4 and 5. As pointed out R^2 values are all very low and if all the dots were similarly colored, I think most readers would agree that there is no obvious difference in metric response at the AWQC. In fact, most of the plots suggest that metrics might be really lower ³ 3-4X the AWQC or the TEL. Given the data shown in these figures, I don't understand how significant differences were observed between affected and unaffected groups in Figures 2 and 3.

Some more specific comments are as follows:

- Line 5, p. 22 - What does "based on the hydrogeochemistry" mean? Be nice to see some summary of the pH and DOC ranges observed.
- Line 9, p. 22 - To what extent was growth a useful endpoint in these analyses? Were effects on survival mostly? Only?
- Line 21, p. 22 - "stressor gradient"? What stressor gradient? You only have affected and unaffected categories.
- Table 4 - Why would you expect relationships between water and sediment measures? They are separate compartments. Not sure this adds much.
- Lines 7-11, p. 25 - Population recruitment fails only if there is no immigration, adaptation, or acclimation of species and the pollutant is persistent or organisms are continually exposed.
- Lines 14-16, p. 25 - This sentence is inconsistent with the logic presented above and suggests a severe disturbance or stress. Some of your data indicates just that.
- Figure 2 - caption should include fish.

- Figure 3 - caption should be just macroinvertebrates.
- Line 2, p. 20 - Should be Table 1 instead of Figure 3.
- Lines 3-4, p. 28 - What was the fish species diversity observed?
- Lines 14-17, p. 28 - So how do you know macroinvertebrates were responding to metals then?
- Lines 1-2, p. 31 - Isn't the bioassay taking into account these factors?
- Lines 17-19, p. 31 - Given this, why were acute tests used in the first place? Also, why didn't you look at acute WQC exceedences in your analyses?

7. *Does the discussion section of this chapter bring out the most important insights of the analyses?*

There is no real discussion section presented. However, in the Results section, the discussion presented brings out most of the insights. However, much of what is brought out regarding the metric results is fairly basic ecology and not really based on data from this study.

8. *Does this chapter line up with the objectives that were stated in the Introduction and were those objectives accomplished?*

Overall, yes. The chapter presents analyses mentioned in the Introduction and generally accomplishes objectives presented, given the issues I brought up under #6.

Virginia Province Estuaries EMAP Study

9. *Is this study presented clearly, and would it be easy to understand for a reader unfamiliar with the three study methods?*

As noted in the answer to question #5 above, the Introduction section is redundant and unnecessary if this is to be a report. Instead, I would suggest starting with the text in Section 3.2.1 and label that Introduction for this chapter.

I do not think the study results are presented clearly or discussed accurately and a reader not familiar with the statistics could not possibly decipher much of the regression results. Table 7 and Figure 7-9 would be unintelligible to most readers and, as noted below, I think the entire regression analysis approach used is a poor way to examine and present these data.

10. *Are the findings and the limitations of the analyses correctly stated?*

No. I have several concerns with the way findings are presented and the conclusions drawn. First, the authors should rethink the ANCOVA approach using % silt/clay. While I agree that substrate size can affect bioassay results and contaminant bioavailability, the reader has no idea how variable

% silt/clay really is across the 201 sites used in analyses. Given that the authors limited analyses to poly-eurohaline sites only (a wise move), I would expect particle size to be relatively fine at all sites. If not, couldn't the authors limit analyses to a certain range of % silt/clay and thereby avoid the entire issue? The authors need to present analyses demonstrating the effect of % silt/clay on measurement endpoints to see if it's really a significant factor. What do the results look like if % silt/clay is not included in the analysis? After all, the authors don't break out TOC gradients, yet that is at least as important as % silt/clay.

Second, the Field et al. model relied in part on the data used in this study. The authors need to demonstrate why it is acceptable to use a model calibrated, in part, with these data, as a means to examine chemical-bioassay relationships in this study.

Third, I have the same concerns with the y analysis here as I noted in answer #5 above. Most of the sites are unaffected apparently.

Fourth, the chapter does not present any real data to support the general conclusion on lines 16-19, p. 55. This study found little in the way of organism effects.

Fifth, given the limitations of the analyses mentioned on pp. 59-64, I fail to see how the authors can claim to have seen any relationships of significance. These sediments were likely affected by stressors other than metals (e.g., PAHs in the Elizabeth River) as most of the analyses show. 10-d sediment tests won't pick up less acutely toxic conditions, leaving one to scratch their head as to why a particular metric gives the results it does.

Specific comments follow:

- Figure 6 - How do you have >100% survival?
- Line 12, p. 53 - Are you sure Field et al. used 90% survival to classify toxic sediments?
- What is "percent composition metrics", line 5, p. 55?
- The regression lines in Figures 7-9 don't jive with the dots in many cases nor do the 95% C.I. How were these regression lines computed?
- How is TOC in the sediment a stressor? (line 4, p. 62).
- Lines 7-9, p. 62: another explanation is that the 0.5 threshold in the model is wrong.
- Line 22, p. 62 - line 2, p. 63: This is stated incorrectly. If SEM/AVS is < 1.0, then metals can not cause toxicity. If the ratio is > 1.0, you don't know whether it's toxic or not.
- Lines 21-23, p. 63: You would also expect spurious differences given the large difference in sample size between groups.

- Lines 1-2, p. 64 seem to me to be directly antithetical to the statement made in the next sentence. The fact is, the different measures often disagree at many sites.

11. *Does the discussion section of this chapter bring out the most important insights of the analyses?*

    There is not real discussion section per se but the discussion presented brings out most of the important "insights". However, I think the conclusions are generally unsupported by the data and that the regression analyses show little or no relationship between measures.

12. *Does this chapter line up with the objectives that were stated in the Introduction and were those objectives accomplished?*

    Yes, given the caveats noted above and also in answer #8 above.

Conclusions

13. *Are the conclusions stated in this chapter correct (according to your understanding of the problem outlined in the Introduction)?*

    Not really. The first sentence needs caveats - relationships were observed with only a few metrics and over the entire dataset. Disagreement among measures at a large proportion of the sites indicates little or no real relationship among measures in these studies.

    The discussion concerning metrics and lack of stressor-specificity should acknowledge the recent work done in this regard.

    The statement that chronic measures should be more predictive of community-level effects is theoretical at best. This study presents no data either way on this point.

    The statement concerning the policy of independent application seems to come out of left field here. Nothing in this entire report really addresses this policy in a direct way. In fact, all of the discussion in the two studies, and prior to this sentence in the Conclusion gives the reader the impression that the community metric results are the final arbiter of effect. All of your analyses were designed to test whether either chemistry or bioassay thresholds agreed with the community results.

    The last paragraph of the Conclusions does not follow exactly from everything before it. You went to great lengths to discuss how both ambient bioassays and chemical analyses can misinform one about the true condition. Your study just showed that the biology would be apparently impaired yet bioassays and chemistry tell you otherwise. The strength of evidence analysis idea needs far more discussion in light of the difficulties brought out in these studies.

14. *Are the conclusions correctly derived from the information presented in this document, and does the text of this chapter appropriately refer to those findings and adequately support the conclusions?*

    The conclusions seem fairly disconnected from the findings. There is little or no reference to findings in the Conclusion section.

15. *Are there any other conclusions that can be derived from the findings reported in this document that should be added to those presented?*

    Yes. Need to distinguish between general regional patterns or relationships from those that pertain to any given site. Classification of "impairment" using chemical or bioassay thresholds is an art not a science and the thresholds typically have tremendous uncertainty. If you're going to use biological metrics (which I think is fine) they need to be calibrated for the region and based on some knowledge of responsiveness to stressors of concern. The shotgun approach to metrics used here is not very useful. Other comments presented earlier provide other conclusions that I think should be presented here.

16. *Does this chapter line up with the objectives that were stated in the Introduction and were those objectives accomplished?*

    The Conclusion seems to be somewhat removed from the rest of the report. Most of the Conclusion is a rehash of the Introduction in terms of how the various measures differ in their level of biological organization. I'm not sure we're any closer in terms of how those levels relate to each other.

Introduction Revisited (re-read the *Introduction* following your review of the document)

17. *Having read the document, would you say its nature, purpose and limitations are accurately described in the Introduction?*

    Not exactly. The entire issue of <u>multiple stressors</u> is completely avoided in the Intro yet that is a large source of the difficulties encountered according to the authors. Also, the use or misuse of independent application appears to be unaddressed in the studies. The study is about relationships only, not which measure yields artifacts or true results.

Executive Summary

18. *Is the Executive Summary easy to read?*

    I don't see an Executive Summary - only an Abstract. It is easy to read.

19. *Will a reader who does not examine the rest of the document get an accurate view of its contents, key findings, and its limitations from reading the Executive Summary alone?*

    Not really. The large disagreement among measures at a substantial proportion of the sites is not mentioned. Nor is the degree to which very polluted sites in the dataset drove the relationships observed. The many caveats discussed in the report (e.g., acute bioassays versus chronic chemical thresholds) are not even hinted at here.

General Comments

20. *Please state your overall assessment of the technical quality and scientific accuracy of this document, and provide any suggested changes needed.*

    Please see the many comments above. The technical quality is poor to mediocre in my opinion and the study was not well-designed to address the objectives. Even under the best of circumstances, such field-lab studies are very complex and difficult to decipher. There needs to be a better job selecting appropriate and related measures (e.g., chronic bioassays with chronic chemical thresholds). This is particularly so for the community data: calibrated metrics, known reference site/condition data, and probably multivariate approaches should also be used to address study objectives.

Literature Cited

Diamond, J**.** and C. Daley. 2000. What is the relationship between whole effluent toxicity and instream biological condition? *Environmental Toxicology and Chemistry*19:158-168.

Groethe, D.R., K.L. Dickson, and D.K. Reed (eds). 1996. Whole Effluent Toxicity Testing: An Evaluation of Methods and Predictability of Receiving System Responses. SETAC Publications, Pensacola FL.

REVIEW COMMENTS OF REVIEWER 2

Thomas W. La Point, Ph.D.
Professor and Director, Institute of Applied Sciences
University of North Texas
1704 West mulberry
Denton, TX 76203
940-369-7776
Email: lapoint@unt.edu

Introduction

1. *Does the introductory chapter make a coherent statement about the nature, purpose and limitations of this document, and of the research it describes?*

   Yes, it does. I would recommend, however, that the report be revised to appear less of a combination of two papers. There is much repetition in the text that could be dropped with no loss of explanatory power.

2. *For those areas within your expertise, is the information accurate, clear and concise?*

   Yes.

3. *For those areas outside your expertise, is the information clear, concise and easy to follow?*

   Yes.

4. *In terms of completeness, organization and level of detail, does the information seem to provide an appropriate introduction to the topics covered, for the purposes of this document.*

   I think the value of the report could be enhanced. I would recommend expanding the introduction to discuss, once, the use of segmented regression. Rather than having it described twice in the two chapters, it could be a section in "Analytical Approaches to Linking Field Responses to Measured Chemical or Toxicological Endpoints."

   Some editorial suggestions:

   - *Page 2; line 10: azteca is misspelled.*
     The misspelling of the specific epithet, *azteca*, has been corrected.

   - *Page 2; line 16: such is misspelled.*
     The space has been removed from the word, such.

   - *Page 3; line 14: "while" implies duration of time. I would recommend replacing "while" with "because they"*
     We have made the suggested change.

   - *Page 3; line 15: delete "usually," and change "working" to "work."*
     We have deleted "usually", but changed the rest of the phrase to "quantify characteristics of selected biotic assemblages". This also achieves the effect suggested by the reviewer.

Southern Rocky Mountain Ecoregion R-EMAP Study

5. *Is this study presented clearly, and would it be easy to understand for a reader unfamiliar with the three study methods?*

   The toxicity bioassays, the sampling regime for benthos, and the chemical analyses were adequate and appear to fully describe the techniques used. The fish sampling did not relate sufficient detail concerning time and distances that were electro-shocked. If this report is to be used as a stand-alone document, it may behoove the authors to expand on the electro-shocking technique as folks may not be as familiar with it.

   It is here, in Section 2.1., that there is much text that was stated in the Introduction and gets re-stated in the next chapter also. Specifically, I would place the text on pages 8 and 9 into a broad introduction for both freshwater and estuarine examples and not have those sections repeated in each of the chapters.

   - *Page 10; line 3: this is no longer a "paper," but a chapter in a report.* The suggested change has been made.

   - Page 10; line 6: delete "are" and change "predictive of" to "predict." The suggested change has been made.

6. *Are the findings and the limitations of the analyses correctly stated?*

   The title of Table 1 is very confusing. The title should be more explanatory that the table is not actually "macroinvertebrate and fish metrics…", but rather the table is comprised of F-values for comparisions of affected and non-affected sites, with levels of significance. The F-values are categorized by community metrics and by measured endpoint, e.g., dissolved metal criterion, 48-hr C. dubia bioassay, sediment threshold values for Hyalella, and 7-d sediment Hyalella bioassays.

   Page 18; line 16ff: I strongly recommend this section, on the index of correspondence, be included in an expanded introduction at the beginning of the report. It is used in both chapters (FW and estuarine), so it should be stated once and expanded. It is a good index to use. It just needs some highlighting. My recommendation stems from my assumption that the report will be used as a stand-alone report and provide guidance to assessors.

   The segmented regression described on page 20 ff is highly useful and one of the best components of this report. I strongly suggest bringing it up into the expanded introduction and expanding on its use. The value of Figure 4 is high! That type of analytical approach may have the best opportunity to be used in other examples of sites influenced by contaminants.

7. *Does the discussion section of this chapter bring out the most important insights of the analyses?*

   Yes, it does. If the report is to be used as a guidance document for regional or state assessors, there could be more text allocated to explaining the segmented regression approach, as I stated above. There could be a better explanation of the index of correspondence. For example, does the difference between a value of 0.89 and 0.83 merit the conclusion (paraphrasing from Page 22; lines 3ff), "the index was slightly greater for the association between water-based assessments than sediment-based." What is the "power of the test" for this index?

   - *Page 22; line 2: replace "while" with "although."*
     We have replaced "while" with "although" or "whereas" as appropriate.

   - *Page 22; line 3: insert "correspondence among groups," between The and n. Delete "index".*
     We have changed the sentence as suggested.

   - *Page 22; line 4: insert "for" between …"than" and "those."*
     We have changed the sentence as suggested.

   - *Page 22; line 12: replace "while" with "whereas."*
     Line 13: replace "while" with "although."
     We have replaced "while" with "although" or "whereas" as appropriate.

   - *Page 32; discussion on the segmented regression: Excellent! The information depicted in Figures 4 and 5 will be the most highly cited components of this report (in this reviewer's opinion).*
     No changes required.

   - *Page 35: line 8: replace "while" with "because."*
     Changing "while" to "because" in this sentence is not appropriate and would change the meaning of the sentence.

   - *Page 35; line 12: delete "the" between "into" and "unaffected."*
     We have changed the sentence as suggested.

8. *Does this chapter line up with the objectives that were stated in the Introduction and were those objectives accomplished?*

   Yes.

Virginia Province Estuaries EMAP Study

9. *Is this study presented clearly, and would it be easy to understand for a reader unfamiliar with the three study methods?*

   This chapter has several duplications (e.g., see sections on Page 37' lines 18ff; page 38, lines 16ff; page 39, lines 3ff). These sections should be pulled into one comprehensive introduction, as the concepts covered in how the metrics work, the use of the correspondence analysis, and the general sampling techniques are similar.

   - *Page 37; line 5: Delete "The purpose of" and begin the sentence with "This chapter." Delete "was to" and replace "compare" with "compares."*

   We have rewritten Section 3.1 in response to other comments by the external peer reviewers, and this editorial suggestion no longer applies.

   - *Page 39; lines 7-8: Begin this sentence with "This chapter applies our approach to the effects…".*
     We have rewritten Section 3.1 in response to other comments by the external peer reviewers, and this editorial suggestion no longer applies.

   - *Page 40; line 7: replace "while" with "and."*
     We have changed the sentence as suggested.

   - Line 12: delete "generally," The samples were collected with a Van Veen grab (or not, and if not, with what? That should be described).

   - Line 14: Delete "The" and begin the sentence with "Samples…"

   - Line 16: Delete "The" and begin the sentence with "Organisms…"
     We have changed the sentence in line 7 as suggested. The detail in Line 12 is that in a few cases, only 2 grabs were collected. We have changed the sentence to clarify this. We have deleted "The" from the beginning of the two sentences as suggested.

10. *Are the findings and the limitations of the analyses correctly stated?*
    Yes. I do have a question on the extensive chemical analyses described on Page 41. Would it be of value for this report to include either an http web site or a CD with the chemical methods? Depending on how the report is to be used, such an appendix might enhance the overall utility of the report.

11. *Does the discussion section of this chapter bring out the most important insights of the analyses?*

Yes. The discussion of the index of correspondence on page 48 (lines 3-6) duplicates what has already been written in Chapter 2. I would delete this section and place into a more comprehensive introduction.

- *Page 49; line 19: delete "while" and replace with "whereas."*
  Line 19: insert "did" between "than" and "the."
  Line 20: replace "while" with "and."

  We have changed the sentence as suggested, except we replaced "while" with "whereas" in both cases.

12. *Does this chapter line up with the objectives that were stated in the Introduction and were those objectives accomplished?*

Yes.

Conclusions

13. *Are the conclusions stated in this chapter correct (according to your understanding of the problem outlined in the Introduction)?*

The conclusions follow from the discussions in the two chapters. However, in my opinion, the conclusions do not go far enough. There should be a re-statement of the importance of the segmented regression in quantifying logistic data (e.g., those based on thresholds or a criterion or test response, etc). It is very powerful and needs to be re-visited in this section.

14. *Are the conclusions correctly derived from the information presented in this document, and does the text of this chapter appropriately refer to those findings and adequately support the conclusions?*

The conclusions are too weakly stated. There is much more value than stated. One other item that might be included is the fact that such analyses are data intensive and require a large data set.

The only other component that should be discussed somewhere in this report is the "power of the test," or Type II error rate. How much of a difference among sites or among responses is needed to determine that "they are or are not affected?" A discussion of this would lend itself to the useful suggestion that assessors think about sample size in the conduct of such assessments.

15. *Are there any other conclusions that can be derived from the findings reported in this document that should be added to those presented?*

As I stated above, the importance of sample size, the value of segmented regression, and the overall need for establishing a framework with which to implement an assessment is critical and would be of value to emphasize in this report.

16. *Does this chapter line up with the objectives that were stated in the Introduction and were those objectives accomplished?*

It does not address the third objective (page 5; line 3). The level to which these techniques are predictive or protective has not been established. That the techniques work in mining, freshwater systems is demonstrated. Without a discussion of the power of the test, we do not know the sensitivity of the techniques in further cases.

Introduction Revisited (re-read the *Introduction* following your review of the document)

17. *Having read the document, would you say its nature, purpose and limitations are accurately described in the Introduction?*

Yes, but with the caveat that the Introduction should be expanded to include a larger discussion of the regression and correspondence analyses. These sections could then be cut from Chapters 1 and 2, saving text and space. Generally, however, the introduction does state what is to be analyzed and how, and for what purposes. It does lead us into the report.

Executive Summary

18. Is the *Executive Summary* easy to read?

There is no Executive Summary in the version I received. Do you mean the Abstract? If so, the abstract is generally adequate, but could be enhanced by mentioning the use of regression techniques.

19. *Will a reader who does not examine the rest of the document get an accurate view of its contents, key findings, and its limitations from reading the Executive Summary alone?*

No.

General Comments

20. *Please state your overall assessment of the technical quality and scientific accuracy of this document, and provide any suggested changes needed.*

Overall, I very much like the information presented in this report. I would encourage editing the report so it does not appear as much like two separate papers put together. The methods and analyses sections should be combined into one, where there are similarities. Where there are differences in approach to FW or estuarine situations, these should be dealth with in the appropriate chapter.

I would certainly make more of the regression approaches (and limitations of regression).

REVIEW COMMENTS OF REVIEWER 3

Gary M. Rand, Ph.D.
Professor, Department of Environmental Studies
Southeast Environmental Research Center (SERC)
Florida International University
3000 N.E. 151st Street
North Miami, FL 33181
305-919-5869
Email: randg@fiu.edu

1. The introductory chapter is coherent in its statement. However, it does not discuss the limitations in detail. There are several things that could be expanded upon-
   - Explain/define ER-M and PEL/TEL in more detail
   - Summarize a “few” studies on the use of bioassessments in aquatic toxicology
   - Also change the word “bioassay” throughout the text to toxicity test(s)
   - Explain “ambient toxicity tests” in more detail
   - The authors should summarize other studies in the literature that followed a similar approach in the two case studies

2/3. The information appears clear and concise.

4. I do think the authors should provide an overview of the statistical procedures to be used in analysis. The study is presented clearly but the limitations of each method should be discussed as well.

5. The calculation of the community metrics should be discussed in more detail. Once again, data handling and analysis needs more detail.

6/7. The Tables & Figures could be explained in greater detail. Findings should be explained “more clearly.” Especially 2.3.2 (p. 22) and 2.3.3. (p. 31).

8. The chapter does “line up” with the objectives and they were accomplished.

9. The study is presented clearly.

10. The limitations of the analysis should be discussed in more detail. Once again, data handling & analysis should be explained in greater detail.

11. The most important points of the analyses are discussed. The limitations (p. 63) are relevant to both case studies. The latter should be incorporated into the Introduction.

12. The chapter does “line up” with the objectives and they were accomplished.

13-16. The conclusions are correct but also should be explained more clearly.

17. The nature and purpose are accurately described but the limitations are not accurately described in the Introduction.

18. The Executive Summary is easy to read.

19. In my opinion, a reader will <u>not</u> get an accurate view of the <u>key findings & limitations</u> from reading the Executive Summary (Abstract) alone. Additional and more specific results should be provided.

20. The technical quality and scientific accuracy were fine. The major issue associated with the document is that consideration should be given to a more thorough discussion of certain topics that were explained above.

SECTION II

COMMENT DISPOSITION REPORT

Comment Disposition Report

Reviewer 1: Jerome Diamond
Reviewer 2: Thomas La Point
Reviewer 3: Gary Rand

Introduction

Question 1: Does the introductory chapter make a coherent statement about the nature, purpose and limitations of this document, and of the research it describes?

Comment: *Presents the nature and purpose fairly well but not the limitations. There is no material presented that would tell the reader how generalized the approach or results of this study are. Also, no real limitations presented in terms of the type of data available and data characteristics. (Reviewer 1)*

Response: The following paragraph, which discusses the limitations, has been added to Section 1.1:

Several limitations are imposed on our assessment by use of these data sets and by technical aspects of the three methods used for the ecological assessment of contaminant exposure and effects. These data sets were collected for purposes that were different from those for which they are used in this report. As a result, some aspects of their study design are not optimal for our purposes. For example, the ambient toxicity tests conducted in both studies were acute in duration (EPA, 1993; 1994a; 1994b), whereas the results of chronic toxicity tests would have been more comparable to the community metrics, which generally reflect longer-term effects. Also, technical differences among the three methods go beyond the methods' differences in the levels of biological organization used as their measurement endpoints. For example, differences are related to laboratory testing versus field sampling and the selection of test species that are amenable to their use in a laboratory setting. The intent of this report is to address the relationships among the measurement endpoints used by the three methods. However, these aspects of study design and technical differences among the methods are discussed in the following chapters to clarify how they affect the observed relationships among the measurement endpoints.

Comment: *Yes, it does. I would recommend, however, that the report be revised to appear less of a combination of two papers. There is much repetition in the text that could be dropped with no loss of explanatory power. (Reviewer 2)*

Response: We have revised the Chapter 1 to concentrate the introductory information repeated in the introductions to Chapters 2 and 3 in Chapter 1. We then revised the introductions to Chapters 2 and 3 to remove this repeated information.

Comment: *The introductory chapter is coherent in its statement. However, it does not discuss the limitations in detail. There are several things that could be expanded upon-*

- *Explain/define ER-M and PEL/TEL in more detail*
- *Summarize a "few" studies on the use of bioassessments in aquatic toxicology*
- *Also change the word "bioassay" throughout the text to toxicity test(s)*
- *Explain "ambient toxicity tests" in more detail*
- *The authors should summarize other studies in the literature that followed a similar approach in the two case studies (Reviewer 3)*

Response:

- We have defined ER-M and PEL in more detail.
- We have changed the word, bioassay, to the phrases, toxicity test(s) or ambient toxicity test(s) throughout where appropriate.
- We have explained ambient toxicity tests in more detail.
- We have added several paragraphs to Chapter 1 that summarize each study mentioned as having done similar comparisons of toxicity tests and community surveys as follows:

Mount et al. (1984) and related studies compared the results of chronic 7-day tests with Ceriodaphnia spp. and P. promelas of serial dilutions of effluents and of ambient water and the results of community surveys of fish or macroinvertebrates. Their study reaches included from one to more than ten point sources, which included publically-owned treatment plants (POTWs), industrial plants, and chemical plants. Community measurements included the total number of taxa, total density, Shannon-Weaver species diversity, a community-loss index, and the density and percentage composition of individual species and of major taxa, such as Ephemeroptera, Trichoptera, Chironomidae, and Mollusca.

Birge et al. (1989) compared the results of 8-day embryo-larval tests with P. promelas of ambient water and the results of community surveys of macroinvertebrates and fish. Their study reaches were upstream and downstream from a POTW, and community measurements included Shannon-Weaver species diversity, a coefficient of dominance, species richness, total density, the percent composition of macroinvertebrate functional groups, and the presence or absence of fish species.

Eagleson et al. (1990) compared the results of chronic, 7-day tests with C. dubia of effluents taking into account the site-specific dilution of the effluent in the receiving stream and the results of community surveys of macroinvertebrates conducted upstream and downstream of the effluent discharge. The sources of the effluents were classified as either municipal or industrial. Community measurements were total taxa richness and the taxa richness of major taxa groups, such as Ephemeroptera, Plecoptera, Trichoptera, Chironomidae, Oligochaeta, and Crustacea.

Dickson et al. (1992) reanalyzed data from several of the above studies along with data from the Trinity River collected upstream and downstream six major POTWs. The Trinity River study compared short-term, chronic tests with C. dubia and P. promelas of ambient water with the results of community surveys of macroinvertebrates and fish. Community measurements were fish or macroinvertebrate richness and evenness, and a fish index of biotic integrity.

Clements and Kiffney (1994) compared the results of chronic, 7-day tests with C. dubia of ambient water collected along a metal contamination gradient upstream and downstream of California Gulch, a point source of mine drainage to the Arkansas River, with the results of community surveys of macroinvertebrates. Community measurements were taxa richness, total abundance, and the percent abundance of Ephemeroptera and Orthocladiinae.

Question 2: For those areas within your expertise, is the information accurate, clear and concise?

Comment: *This chapter and others need technical editing in many places. Some text is awkwardly constructed and difficult to follow (e.g., lines 9-12, p. 3). Specific technical suggestions are:*

(a) The EPA tox test references (p. 2) are old. If these are what were used in the study (as opposed to the higher QA/QC stipulated in more recent methods), then I don't think the authors can claim use of current methods

(b) Sediment toxicity effects are usually predetermined by comparing test sites to reference sites (see ASTM standards on this) not controls (line 17, p. 2). Controls are used for QA/QC purposes.

(c) Metrics are one tool used to analyze bioassessment data - not the only tool as one would think reading the last paragraph p. 2. Multivariate tools are also used extensively (e.g., RIVPACS).

(d) Lines 3-5, p. 3 are vague. It is not clear from this how metrics are to be used.

(e) I disagree that chemical criteria are organism-based. They are calculated based on toxicity data but the criteria are meant to represent populations not organisms - hence the difficulty with using criteria for endangered species protection (which is organism-based).

(f) References cited in lines 3-5, p. 4 are old. Should reference the Pellston SETAC book on this topic (Groethe et al., 1996) at least. Others are included in a special ET&C issue on WET (e.g., Diamond et al., 2000).

(g) The use or lack of use of <u>many</u> community metrics is not necessarily an issue. The metrics need to be calibrated for the site and it is important that redundant metrics are removed. (Reviewer 1)

Response: (a) The EMAP field surveys that were used in the two analyses were conducted before the most recent editions of these guidance documents were published. Therefore, citing these older editions is appropriate, particularly in the Materials and Methods sections of Chapters 2 and 3. Our understanding of the changes in these bioassays between the two editions is that they are minor and would not affect the conclusions based on these bioassay data. However, as the reviewer's main point is the use of "current methods", we have removed the words "current" or "currently" from any sentences that refer to the bioassay methods.

(b) The reviewer's comment misstates how a significant reduction in growth or survival is determined in sediment bioassays. Survival and growth in the test bioassay are compared with that in a concurrently run negative control. Moreover, we are describing what was done in the two EMAP field surveys. Some deviations from standard methods as described in documents, such as the ASTM standards, may occur, but as all samples within each survey were treated in the same way, this would not affect our results.

(c) The context of this discussion is the approach to analysis of bioassessment data used by the USEPA in contrast to the other methods (i.e., chemical criteria and ambient toxicity tests). While multivariate tools, such as RIVPACs, are used elsewhere and current interest exists for incorporating such tools into the USEPA's approach, we feel discussing such tools would confuse the point of the document and is not appropriate in this introduction.

(d) We have rewritten the two sentences and added references that further discuss these two concepts to clarify how the metrics will be used in this document.

(e) This is one of the misconceptions about toxicity tests. As discussed in USEPA (2003d), the generic ecological assessment endpoints document, the measurement endpoints of bioassays, such mortality, growth, and fecundity, are organism-level endpoints, while population-level endpoints are population size and rates of population change. Such population-level endpoints may be extrapolated from the organism-level endpoints, if one makes certain assumptions. The numerical methods used to estimate chemical criteria extrapolate the organism-level bioassay data to a number that is protective at the community-level, by setting a criterion below a concentration that causes adverse effects to a small proportion (i.e., usually 5%) of the tested species. This is discussed in the introduction. The difficulty with using criteria to protect endangered species is that toxicity tests are almost never conducted with endangered species, and as a result, uncertainty exists whether an endangered species is in that small proportion of the community that is theoretically not protected by a criterion. Therefore, no changes are required in response to this comment.

(f) After re-reviewing Grothe et al. (1996) and the 2000 annual review issue of *Environmental Toxicology and Chemistry* that included Diamond and Daley (2000), we have concluded that most of the literature described in these documents discusses whole effluent toxicity (WET) tests as opposed to ambient toxicity tests. These two types of tests are very similar in methodology, because both use laboratory bioassays with standard organisms to test environmental samples. However, WET tests assess the toxicity of whole effluents, whereas ambient toxicity tests assess the toxicity of contaminants diluted in receiving waters. For the results of a WET test to be extrapolated to effects in an ecosystem receiving the tested effluent, one must assess the change in toxicity because of dilution and other changes resulting from the mixing of the effluent with receiving waters. The results of an ambient toxicity test are more closely related to effects in the contaminated ecosystem, because the changes in toxicity have generally occurred before the sample is collected. When considering studies that have tried to compare the results of ambient toxicity tests to effects in contaminated ecosystems, these two documents cite practically the same literature as does our report. Therefore, we have just added a citation of Diamond and Daley (2000) to our report.

(g) Many of the early studies cited here, particularly Mount et al. (1984) and the related studies, use a few of what are now considered metrics to quantify biotic communities in the streams they studied. In particular, these metrics include taxa richness, the Shannon - Weaver diversity index, a community loss index, and mean density. Today, many additional metrics that quantify different characteristics of biotic communities have been proposed. Most often, individual metrics are calibrated for a region, such as an ecoregion, where the metrics are expected to vary similarly in order to normalize those metrics to a particular range (e.g., 0-10) before combining them into an index of biotic integrity. All the sites in the Colorado REMAP study were in a single ecoregion, the Southern Rockies, so there is no need to calibrate the individual metrics, particularly since we tested them individually. Moreover, since the metrics were tested individually, there is no need to exclude redundant metrics. This is only needed when combining the metrics into an Index of Biotic Integrity, which we did not do in this report. Because even metrics like Ephemeroptera richness, Ephemeroptera and Plecoptera richness, or EPT richness are not completely redundant, these types of analyses, where individual metrics are compared to a single stressor, allow us to assess which metric is most sensitive to the stressor of interest.

Comment: *Yes. (Reviewer 2)*

Response: No response is required.

Comment: *The information appears clear and concise. (Reviewer 3)*

Response: No response is required.

Question 3: For those areas outside your expertise, is the information clear, concise and easy to follow?

Comment: *All information was within my area of expertise. (Reviewer 1)*

Response: No response is required.

Comment: *Yes. (Reviewer 2)*

Response: No response is required.

Comment: *The information appears clear and concise. (Reviewer 3)*

Response: No response is required.

Question 4: In terms of completeness, organization and level of detail, does the information seem to provide an appropriate introduction to the topics covered, for the purposes of this document?

Comment: *Level of detail is appropriate. Seems complete. (Reviewer 1)*

Response: No response is required.

Comment: *I think the value of the report could be enhanced. I would recommend expanding the introduction to discuss, once, the use of segmented* *regression. Rather than having it described twice in the two chapters, it could be a section in "Analytical Approaches to Linking Field Responses to Measured Chemical or Toxicological Endpoints." (Reviewer 2)*

Response: Unfortunately, segmented regression was not used in the analyses of both surveys. It was only used in the analyses of the survey of Colorado streams. Segmented regression is a technique for identifying thresholds, which ambient water quality criteria and threshold effects levels attempt to identify. On the other hand, the logistic regression approach of Field et al. (2002) does not identify a threshold. Therefore, we decided that using segmented regression in the analyses of the survey of Virginian estuaries was not appropriate, and instead, we used multiple regression. Because segmented regression is not used in the analyses of both surveys, we do not believe that highlighting the method in the Introduction is appropriate. This would give the impression that segmented regression was used in the analyses of both surveys and would likely confuse readers.

Comment: *I do think the authors should provide an overview of the statistical procedures to be used in analysis. The study is presented clearly but the limitations of each method should be discussed as well. (Reviewer 3)*

Response: As discussed elsewhere in response to similar comments by the other reviewers, there are similarities and differences between the statistical procedures used in each analysis, and to combine these statistical procedures into a section in the Introduction would likely confuse the reader about which procedures were used in which study.
Also, we feel that a discussion of the limitations of each assessment method is best left in the Results and Discussion section of Chapters 2 and 3, because these limitations explain many of the cases where the methods disagree.

<u>Southern Rocky Mountain Ecoregion R-EMAP Study</u>

Question 5: Is this study presented clearly, and would it be easy to understand for a reader unfamiliar with the three study methods?

Comment: *I realize the chapters were prepared as somewhat stand alone chapters, but if this is to be a report, there is no need to repeat the introduction given in Chapter 1. Pages 8 and 9 should be deleted. Start with first real paragraph p. 10.*

The study is presented fairly clearly and much of it would be understood by environmental scientists but probably not by a non-biologist. There are several things that I think should be clarified:

(a) Figure 1 - include a scale so someone understands what size area was sampled.

(b) *Were there any other potential sources of stress for the seven "downstream" sites besides mining?*

(c) *How did flows compare between 1994 and 1995? Demonstrate to the reader that combining the data from both years in one analysis is scientifically valid.*

(d) *What was the variability between visits to a given site (line 3, p. 12)?*

(e) *Should include table with method used and detection limits achieved for each metal analyzed.*

(f) *What QC tests were conducted to support the toxicity tests? Reference toxicant tests?*

(g) *Should include acceptability criteria used for each type of toxicity test.*

(h) *The Hyalella test appears to be a non-standard test. Normally this is a 10-d test, not a 7-d test. Why was a 7-d test used? Growth is difficult enough to measure in a 10-d test, I find it difficult to believe that 90% growth of controls could be significantly different. The "preliminary comparisons" (line 21, p. 13) need more explanation and some back-up because much of the analyses depended on the thresholds determined from this comparison.*

(i) *Were the sediment samples for chemistry and toxicity taken from the riffle areas at which macroinvertebrates were collected and used in analyses?*

(j) *Tables 1 and 2 appear to be a laundry list of metrics. Given the preponderance of metals in this region as a main stressor, it would seem more fruitful to examine those metrics that have been shown to be responsive to metal stress. I'm concerned that when so many metrics are examined (there's more than 50 represented here), just by chance, you might expect to see significant differences for a few of them. At the very least, I would strongly recommend looking at the correlations among metrics and delete those that are redundant.*

(k) *Table 1 - I don't understand how the "D" column F values were derived when you have four different metals potentially. In general, I found it difficult to interpret the meaning of this table. As noted previously, if there were fewer metrics examined, and they were uncorrelated with each other, that would help the readability of this table.*

(l) *How come the authors didn't look at additivity of metals as one factor, rather than looking at each one individually in correspondence analyses? The latter approach inappropriately lumps together sites where only one metal exceeds a threshold with sites that exceeded multiple thresholds. There is also no accounting for <u>how much greater</u> the concentration at a site is compared to its threshold. This too adds uncertainty to the analysis.*

(m) *Disagreement assessment of metrics (lines 3-12, p. 20) is not clear for metrics that mean "worse" conditions as they increase in value. (Reviewer 1)*

Response: We have rewritten the introductions in each of the first three chapters to remove the repetitive material from Chapters 2 and 3 and consolidate it in Chapter 1. This was also suggested by the other reviewers.

As a technical report targeted for the Office of Water and Office of Solid Waste and Remedial Response, the intended audience is environmental scientists.

(a) We have added a scale to Figure 1.

(b) Previous analyses also identified increased nutrients and fine sediments and decreased canopy cover associated with livestock grazing in riparian zones as another stressor gradient in these Rocky Mountain streams. We discuss this in the results and discussion section for Chapter 2. Comparisons of nutrient concentrations between upstream and downstream pairs do not suggest that nutrients differed between these sites.

(c) The value of EMAP data sets in these analyses is that the probabilistic sampling design randomizes any between year effects among the sampling sites. Moreover, we did not use data from revisits to the same site, which enhances this randomization. Also, the use of the index period, which was the period of the water year when stable base flows occur in these Rocky Mountain streams, and the avoidance of episodic events, minimizes the effects of seasonal and yearly variation. These characteristics of the Colorado data set are already described in Section 2.2.1, Study Area and Survey Design.

(d) As stated in Section 2.2.1, data from only the first visit to a site were considered in these analyses. Therefore, the variability between visits to a single given sample is not a consideration here.

(e) The methods used for analysis of metals in this Colorado study are described in USEPA (1987), *Handbook of Methods for Acid Deposition Studies: Laboratory Analyses for Surface Water Chemistry* as cited in the text. This document describes a single atomic absorption method for metals. Therefore, a table would not add information to the chapter. However, we will add a sentence that describes the detection limits achieved for the four metals of interest.

(f) We have added statements about the QC used in both the water and sediment toxicity tests.

(g) We have added statements about the acceptability criteria used in both the water and sediment toxicity tests.

(h) The described study was conducted early in the development of sediment bioassay methods with *Hyallela azteca*, and the described 7-day test is what was conducted, although now this duration is shorter than what is currently the standard duration. Dr. James Lazorchak, a coauthor for this chapter, was directly involved in the ambient toxicity tests for the Colorado R-EMAP surveys, and no difficulties occurred in measuring growth in these tests. We have have added minimum significant differences (MSD) calculated following Thursby et al. (1997) to support our selection of significant effects in this test.

(i) Sediment samples were collected from depositional areas near each of the nine interior cross-section transects along a reach. The macroinvertebrate samples were also collected at these interior cross-section transects. We have added a statement about where the sediment samples were collected.

(j) Where is a list of metrics shown to be particularly sensitive to metals? Remarkably few studies have investigated the sensitivity of various metrics to specific stressors, like metals or other contaminants. Part of the intent of this study was to identify such metrics for metals. The concern that with so many individual tests one might expect to see significant differences for a few metrics just by chance was addressed by use of the sequential Bonferroni technique to correct *p*. Because each metric was tested separately, exclusion of metrics that are not redundant is no needed. In fact, comparisons among such metrics might reveal small differences among the metrics that make one more sensitive to the stressor of interest than the other.

(k) As stated in the text, sites were classified as affected if the concentration of at least one of the four metals in water exceeded its chronic AWQC or in sediment exceeded its TEL. We have tried to emphasize this by adding a footnote to Table 3 that says “at least one”. However, the reviewer seems to understand this partially based on his next comment.

(l) When looking at comparisons between chemical data and criteria or at the results of ambient toxicity tests, one is usually faced with a yes or no question. One example would be: “Is the concentration of Cu in water greater than the chronic criterion for water?” Another example would be: “Is the growth of *Hyalella azteca* on sediments from a site significantly less than that on a reference sediment?” If the concentration of Cu in water is greater than the chronic criterion, the answer will be “yes”, no matter if Cu is only slightly greater than the criterion or if concentrations of Cd, Cu, Pb, and Zn are all much greater than their individual criterion. Therefore, grouping such sites in an “affected” group is appropriate. This is a categorical question, and the statistics used were appropriate to a categorical question (i.e., ANOVA and the categorical association among groups, λ). On the other hand, assemblage metrics are continuous variables and were difficult to use to divide the sites into groups by themselves, although we ultimately did this using the 95% lower and upper confidence limits. That is why some analyses with metrics were used to further explore the questions about mixtures of metals and whether the metrics decrease as the extent to which metals exceed criteria increase.

(m) In the analyses of the Colorado survey, all the metrics, which exhibited statistically significant differences between the affected and unaffected groups in the ANOVA when p was corrected using the sequential Bonferroni technique, decreased with increasing metals. Therefore, we only had to describe this situation in Chapter 2 and believe this simplifies our description of the methods. We did modify the indicated sentences to make this clear as follows:

If a community metric decreases as a stressor increases, an assessment based on that metric would differ if the metric was "greater than expected" at a site identified as affected based on organism-level effects or if the metric was "less than expected" at a site identified as unaffected based on organism-level effects. In this study, all the statistically significant metrics decreased in the affected group, and we defined community metrics as "greater than expected" when the metrics were greater than the 95% upper confidence limit (UCL) of an affected group and as "less than expected" when the metrics were less than the 95% lower confidence limit (LCL) of the unaffected group.

Comment: *The toxicity bioassays, the sampling regime for benthos, and the chemical analyses were adequate and appear to fully describe the techniques used. The fish sampling did not relate sufficient detail concerning time and distances that were electro-shocked. If this report is to be used as a stand-alone document, it may behoove the authors to expand on the electro-shocking technique as folks may not be as familiar with it.*
It is here, in Section 2.1., that there is much text that was stated in the Introduction and gets re-stated in the next chapter also. Specifically, I would place the text on pages 8 and 9 into a broad introduction for both freshwater and estuarine examples and not have those sections repeated in each of the chapters. (Reviewer 2)

Response: 1. We have added detail concerning the time and distances electrofished. Specifically, we make it clear that the entire stream reach, defined as a length of stream equal to 40 times the mean low-flow, wetted width (minimum of 150 m and maximum of 500 m), was electrofished. Total collection time was not less than 45 minutes and not longer than 3 hours within the defined sampling reach and was divided in proportion to the area of the stream reach within each of the ten intervals among the eleven cross-section transects.

2. In response to this comment and similar comments by the other reviewers, we have consolidated most of the introductory material in Chapter 1, and removed it from the introductions to Chapters 2 and 3.

Comment: *The calculation of the community metrics should be discussed in more detail. Once again, data handling and analysis needs more detail. (Reviewer 3)*

Response: We have added a discussion of how the different types of community metrics, richness metrics, abundance metrics, composition metrics, evenness metrics, trophic guild metrics, and pollution tolerance metrics, are calculated. The discussion also includes examples of the different types of community metrics. In response to this comment and those of the other reviewers, we have added other details to Section 2.2.7, Data handling and analysis, such as the specific computer software used for each statistical analysis. We believe that this section is very explicit in describing our analyses.

Question 6: Are the findings and the limitations of the analyses correctly stated?

Comment: *I disagree with the interpretation of the findings in several places. First, it is unclear how much extremely contaminated sites are driving any relationships or regressions observed. A cumulative probability plot or a box plot of metal concentrations observed at all sites in the dataset should be done for the bioassay data.*

Second, it is clear that most of the sites have low metal concentrations and no toxicity. The y analysis appears to be biased by the agreement in one cell of the matrix in each case. A true accurate assessment would remove those sites for which there are no apparent stressors present and look at the remaining relationships in which at least one measurement endpoint is exceeded. If this were done, the relationships between endpoints are pretty bleak. I would note that this problem is not unique and has been observed in recent studies including WET-bioassessment comparisons (e.g., Diamond et al., 2000).

Third, as the authors acknowledge, there are some large disconnects between the types of endpoints being compared; e.g., sediment thresholds based on 28d Hyalella tests versus results of 7-d Hyalella sediment bioassays. The fact that acute bioassay results were related at all to benthic metrics suggests to me that there must have been at least a few sites that were severely contaminated with metals (not unlikely in this region), and that results from these sites were driving the statistical relationships observed.

Fourth, I concur with the approach taken in the results and discussion, examining the extent to which chemistry and bioassay measures, in relation to their thresholds, agree with the bioassessment metric data. But the results of this approach only partially supports the authors' conclusions in my opinion. The authors point out the large discrepancy in assessment results using bioassays and chemical thresholds and they also point out the many discrepancies observed between assessments at sites based on metrics and other measures. The point is that on a site-by-site basis, results of different measures are often conflicting (i.e., disagree), negating the conclusion that "organism-level effects are predictive of effects at the community level" (line 2-3, p. 36 and lines 2-6, p. iii, Abstract). The report needs to clearly state that the relationships observed are statistical, based on the data in general, and NOT predictive of relationships at any given site.

Fifth, the conclusions presented for the piecemeal regression analysis appears unfounded to me based on the plots in Figures 4 and 5. As pointed out R^2 values are all very low and if all the dots were similarly colored, I think most readers would agree that there is no obvious difference in metric response at the AWQC. In fact, most of the plots suggest that metrics might be really lower ³ 3-4X the AWQC or the TEL. Given the data shown in these figures, I don't understand how significant differences were observed between affected and unaffected groups in Figures 2 and 3.

Some more specific comments are as follows:

- Line 5, p. 22 - What does "based on the hydrogeochemistry" mean? Be nice to see some summary of the pH and DOC ranges observed.
- Line 9, p. 22 - To what extent was growth a useful endpoint in these analyses? Were effects on survival mostly? Only?
- Line 21, p. 22 - "stressor gradient"? What stressor gradient? You only have affected and unaffected categories.
- Table 4 - Why would you expect relationships between water and sediment measures? They are separate compartments. Not sure this adds much.
- Lines 7-11, p. 25 - Population recruitment fails only if there is no immigration, adaptation, or acclimation of species and the pollutant is persistent or organisms are continually exposed.
- Lines 14-16, p. 25 - This sentence is inconsistent with the logic presented above and suggests a severe disturbance or stress. Some of your data indicates just that.
- Figure 2 - caption should include fish.
- Figure 3 - caption should be just macroinvertebrates.
- Line 2, p. 20 - Should be Table 1 instead of Figure 3.
- Lines 3-4, p. 28 - What was the fish species diversity observed?

- Lines 14-17, p. 28 - So how do you know macroinvertebrates were responding to metals then?
- Lines 1-2, p. 31 - Isn't the bioassay taking into account these factors?
- Lines 17-19, p. 31 - Given this, why were acute tests used in the first place? Also, why didn't you look at acute WQC exceedences in your analyses? (Reviewer 1)

Response: *First* - We have added a set of box plots. The first box plot compares the summed ratios of the dissolved concentrations of Cd, Cu, Pb, and Zn in water to their dissolved AWQCs for the groups classified as affected and unaffected based on the water ambient toxicity tests. The second compares the summed ratios of the sediment concentrations of these metals to their TELs for the groups classified as affected and unaffected based on the sediment toxicity tests. The box plots include the mean, upper and lower 95% CL, and individual values for each site. We believe this shows that the data set is not dominated by a few extremely contaminated sites.

Second - The statistic γ is only one measure of the relationships exhibited in the contingency tables. Another is simply direct inspection of the contingency tables, although further details of the contingency tables are not described explicitly in the text. We have added the following discussion of the contingency tables to the text. This text includes a discussion of the new figure added in response to the first comment above.

The mean summed ratios of the dissolved concentrations of the four metals to their chronic AWQCs and the mean summed ratios of the sediment concentrations of the four metals to their TELs were greater at sites classified as affected by the ambient toxicity tests for water and sediment, respectively (Figure 2). However, these two measures agreed in their classification of a site at only 53% of the 19 sites identified as affected by at least one measure for water and only 34% of the 35 sites identified as affected by at least one measure for sediment.

Third - Contrary to the reviewer's assessment and as shown in the figure cited above, the results of the statistical analyses are not being driven by a few severely contaminated sites. We acknowledge some large disconnects between the types of endpoints being compared (e.g., sediment thresholds based on chronic Hyalella tests versus the results of acute Hyalella sediment bioassays) that go beyond the level of biological organization measured by the measurement endpoints. We discuss these differences in terms of the mismatches between the classifications based on chemistry and on ambient toxicity tests and how the differences affect our ability show the relationships at the community level as measured by community metrics. However, none of this negates the relationships we were able to show with the sensitive community metrics.

Fourth - Based on the analyses presented, we believe that evidence is presented for a relationship between the measurement endpoints used by each method. We also believe this relationship is somewhat predictive, but as discussed extensively, many other differences exist among these methods that go beyond the levels of biological organization used as measurement endpoints and effect this predictive ability. We have added a statement to the conclusions of this study to highlight this. In the end, the measurement endpoints are all indicators of effects, are not perfect measures of the effects of metals in these streams, and will not perfectly predict the effects.

Fifth - Particularly for the water data, the plots change distinctly between the sites where the chronic AWQCs were not exceeded and those where the criteria were exceeded. This was verified with the piecewise regression analysis. As described in the methods, the change is verified when particularly the slope but also the intercept changes signficantly at the join point. We agree the regressions did not significantly change for the sediment data, suggesting that the TELs do not represent thresholds for effects. The r^2 is for the entire regression. For field data, such regressions usually have low r^2 to begin with, but in addition, the model predicts no significant relationship between metals and the metrics when metal concentrations were less than the chronic AWQCs. Therefore, while we report the r^2, it may not be the best measure of how good the relationship is.

- Line 5, p. 22 - It is well known from the mine drainage literature, that as pH increases downstream from a mine source, the solubility of metals changes and the metals precipitate out of solution and are deposited to the stream sediments. Therefore, the gradient in metals downstream of a mine source is greater metals in water close to the mine source and decreasing downstream. On the other hand, metals in sediments increase downstream of the mine source within the zone where metal precipitation occurs. Unfortunately, because of problems with measurement of pH in this study, the pH data were considered invalid. DOC concentrations ranged from less than a detection limit of 1.0 mg to 10.8 mg/L. We have added this information to this paragraph.
- Line 9, p. 22 - Out of 105 sites for which there was sediment toxicity data, 80 sites did not exhibit significantly reduced survival or growth. Of the remainder, 13 exhibited reduced growth, 10 exhibited reduced survival, and two exhibited both reduced growth and survival. However, these details are not really pertinent to the discussion here.

- Line 21, p. 22 - It is still a gradient of metals contamination, although we have reduced it to categorical data, and Griffith et al. (2001) looks at metals contamination as a stressor gradient.
- Table 4 - As discussed in the short description of the hydrogeochemistry of mine drainage, the sources and mechanisms that produce metal contamination in the water and sediments of these streams are interrelated and at least some spatial relationship exists between contaminated water and contaminated sediments (i.e., Where the water from a site is contaminated, a greater probability exists that the sediment from the site is also contaminated. This paragraph and tables show that contaminated water and sediment do not always co-occur. Therefore, treating water and sediment separately in the other statistical analyses is appropriate.
- Lines 7-11, p. 25 - Adaptation and acclimation are mechanisms by which a more tolerant population would resist the effects of a toxicant, like metals. Also, metals are a persistent pollutant, particularly when associated with mine drainage. We have changed the text to say that metals contamination is a persistent pollutant and say that more tolerant species might adapt or acclimatize themselves to the toxicant.
- Lines 14-16, p. 25 - We do not see this sentence as inconsistent with the logic presented in the previous sentences. We are talking about the trends in populations of two potential groups of species within the assemblage, sensitive and tolerant. Sensitive species are those that are more affected by a stressor and usually decrease in abundance and are eliminated from the assemblage in the presence of the stressor. Tolerant species are those that may be adapted to the stressor or may acclimatize themselves in some way to the stressor. Sometimes, such tolerant species may increase in response to reduced competition or predation from species eliminated by the stressor. This is the basis of relative abundance metrics. We have changed the sentences to make it clear that we are talking about sensitive species versus tolerant species.
- Figure 2 - The four graphs in Figure 2 and those in Figure 3 were inadvertently switched in the External Peer Review Draft. We have corrected this.
- Figure 3 - See response to last comment.
- Line 2, p. 20 - The significant difference for these two metrics is shown both in Table 1 and Figure 3. We have added Table 1 first in the parentheses.
- Lines 3-4, p. 28 - Maximum total fish species or subspecies richness was 6 and maximum native fish species or subspecies richness was 4. Of those sites with fish, the mean proportion of fish that were trout was 82.7%, and a mean 97.4% of the trout were

not native species or subspecies. We have added this information to the text.

- Lines 14-17, p. 28 - Because, despite the added variability associated with these other stressors, a significant relationship exists between the metrics and metals concentrations. This is the obvious point of this part of the discussion.
- Lines 1-2, p. 31 - We are talking about criteria in this paragraph.
- Lines 17-19, p. 31 - As stated in the second part of the last sentence, "chronic effects would be reflected by the community metrics." We would have used data from chronic toxicity tests, if they would have been available. However, only acute toxicity tests were conducted in this study. To the extent possible, we matched chronic endpoints with chronic endpoints, which is why we used chronic criteria instead of acute criteria. This is initially discussed in the Introduction, where we discuss the limitations of the data sets.

Comment: *The title of Table 1 is very confusing. The title should be more explanatory that the table is not actually "macroinvertebrate and fish metrics…", but rather the table is comprised of F-values for comparisions of affected and non-affected sites, with levels of significance. The F-values are categorized by community metrics and by measured endpoint, e.g., dissolved metal criterion, 48-hr C. dubia bioassay, sediment threshold values for Hyalella, and 7-d sediment Hyalella bioassays.*

Page 18; line 16ff: I strongly recommend this section, on the index of correspondence, be included in an expanded introduction at the beginning of the report. It is used in both chapters (FW and estuarine), so it should be stated once and expanded. It is a good index to use. It just needs some highlighting. My recommendation stems from my assumption that the report will be used as a stand-alone report and provide guidance to assessors.

The segmented regression described on page 20 ff is highly useful and one of the best components of this report. I strongly suggest bringing it up into the expanded introduction and expanding on its use. The value of Figure 4 is high! That type of analytical approach may have the best opportunity to be used in other examples of sites influenced by contaminants. (Reviewer 2)

Response: *Table 1 -* The information sought by the reviewer was originally placed in footnote a. However, based on this comment, this information has been moved into the table caption so that it is more obvious.

Page 18; line 16ff - The focus of this report is the relationships among the assessments based on the three methods and not on the statistical methods used to investigate those relationships. Although it is a stand-alone report, it is not written as or intended to be a guidance document. Because this document is a stand-alone report, we have tried to be reasonably detailed in describing the methods used but feel that if the reader wants minute details, the reader can go to the guidance documents that provide detailed instructions on how to conduct chemical analyses or bioassays. As stated elsewhere, while some statistical methods, like γ, the index of association for categorical data, are used in both studies, other methods, such as segmented regression, are used in only one study. With the Colorado REMAP data, we used one-way ANOVA, whereas with the Virginia Province EMAP data, we used ANCOVA. Segmented regression was used with the Colorado REMAP data, but regular multiple regression was used with the Virginia Province EMAP data. We feel that combining the statistical methods into a single description and highlighting them in the introduction would end up confusing readers as to the statistical methods used in each study. However, we do recognize the potential value of segmented regression for identifying potential thresholds of effect in relationships between various stressors, not just contaminants, and measurement endpoints such as community metrics, hope to pursue this beyond this document.

Comment: *The Tables & Figures could be explained in greater detail. Findings should be explained "more clearly." Especially 2.3.2 (p. 22) and 2.3.3. (p. 31). (Reviewer 3)*

Response: We have rearranged the text in Section 2.3.2 and added more details to explain the tables and figures in greater detail and more clearly. We added some more details in Section 2.3.3 to explain the figures in greater detail, but much of the results discussed in this subsection are most clearly shown by Figures 3 and 4.

Question 7: Does the discussion section of this chapter bring out the most important insights of the analyses?

Comment: *There is no real discussion section presented. However, in the Results section, the discussion presented brings out most of the insights. However, much of what is brought out regarding the metric results is fairly basic ecology and not really based on data from this study. (Reviewer 1)*

Response: We combined the results and discussion because we felt this best tied our results to our conclusions. The intent of this report was to show the relationships between the organism-level measurement endpoints on which criteria or guidelines are based and that ambient toxicity tests measure and the community-level measurement endpoints of community metrics. While working through these comparisons, we recognized many other technical differences among these methods that have little to do with the level of biological organization used as the measurement endpoint. Our discussion of these other technical differences is what the reviewer calls, “fairly basic ecology”. However, these other technical differences are often ignored, and these technical differences are why these methods are complementary.

Comment: *Yes, it does. If the report is to be used as a guidance document for regional or state assessors, there could be more text allocated to explaining the segmented regression approach, as I stated above. There could be a better explanation of the index of correspondence. For example, does the difference between a value of 0.89 and 0.83 merit the conclusion (paraphrasing from Page 22; lines 3ff), “the index was slightly greater for the association between water-based assessments than sediment-based.” What is the “power of the test” for this index?*
Page 32; discussion on the segmented regression: Excellent! The information depicted in Figures 4 and 5 will be the most highly cited components of this report (in this reviewer’s opinion). (Reviewer 2)

Response: The document was not written to be used as a guidance document for regional or state assessors. However, the reviewer’s point on the potential value of segmented or piecewise regression is well taken, and revisiting segmented regression as a method for investigating stressor - response relationships with monitoring data may be appropriate. We found an early EPA document (Hasselblad et al., 1976) that considered the use of segmented regression to develop air pollutant criteria.

Hasselblad, V., J.P. Creson, and W.C. Nelson. 1976. Regression using “hockey stick” functions. EPA-600/1-76-024. U.S. Environmental Protection Agency, Office of Research and Development, Health Effects Research Laboratory, Research Triangle Park, NC.

The index of association, γ, (Goodman and Kruskal, 1972) that we used with the contingency tables is similar the more familiar r^2 from a Pearson correlation analysis, except that the index of association is intended for use with categorical data instead of continuous data. We are not aware of any methods for determining the power of this statistic or even similar statistics, like r^2. Usually statistical power is used in hypothesis testing, where one would like to determine the ability of a test to identify a statistically significant difference among two means that differ by some known amount. However, the reviewer is probably correct in questioning whether saying that the one index value is slightly greater than the other is appropriate. Therefore, we have changed the sentence just to state what the index values were. We have also modified the description of the index in the Methods and Materials to clarify what the index is.

Comment: *The Tables & Figures could be explained in greater detail. Findings should be explained "more clearly." Especially 2.3.2 (p. 22) and 2.3.3. (p. 31). (Reviewer 3)*

Response: This comment repeats the reviewer's comment in response to Question 6. See the changes we made in response to this comment in Question 6.

Question 8: Does this chapter line up with the objectives that were stated in the *Introduction* and were those objectives accomplished?

Comment: *Overall, yes. The chapter presents analyses mentioned in the Introduction and generally accomplishes objectives presented, given the issues I brought up under #6. (Reviewer 1)*

Response: No response required, although a response is presented elsewhere to the issues the reviewer brought up under #6.

Comment: *Yes. (Reviewer 2)*

Response: No response is required.

Comment: *The chapter does "line up" with the objectives and they were accomplished. (Reviewer 3)*

Response: No response is required.

Virginia Province Estuaries EMAP Study

Question 9: Is this study presented clearly, and would it be easy to understand for a reader unfamiliar with the three study methods?

Comment: *As noted in the answer to question #5 above, the Introduction section is redundant and unnecessary if this is to be a report. Instead, I would suggest starting with the text in Section 3.2.1 and label that Introduction for this chapter.*

I do not think the study results are presented clearly or discussed accurately and a reader not familiar with the statistics could not possibly decipher much of the regression results. Table 7 and Figure 7-9 would be unintelligible to most readers and, as noted below, I think the entire regression analysis approach used is a poor way to examine and present these data. (Reviewer 1)

Response: In response to this comment and similar comments by the other reviewers, we have consolidated most of the introductory material in Chapter 1, and removed it from the introductions to Chapters 2 and 3.

Comment: *This chapter has several duplications (e.g., see sections on Page 37' lines 18ff; page 38, lines 16ff; page 39, lines 3ff). These sections should be pulled into one comprehensive introduction, as the concepts covered in how the metrics work, the use of the correspondence analysis, and the general sampling techniques are similar. (Reviewer 2)*

Response: In response to this comment and similar comments by the other reviewers, we have consolidated most of the introductory material in Chapter 1, and removed it from the introductions to Chapters 2 and 3.

While some of the statistical analyses, such as the index of association, γ, are shared between the two analyses, other statistical methods differ. For example, one-way ANOVA was used in Chapter 2, while ANCOVA was used in Chapter 3. Piecewise regression was used in Chapter 2, while multiple regression was used in Chapter 3. Moreover, since the two studies were conducted in different aquatic habitats, streams versus estuaries, the field methods vary as do the community metrics calculated. Also, a logistic regression was used to identify affected sites using sediment chemistry in Chapter 3, while AWQCs and sediment TELs were used in Chapter 2. While some duplication occurs in the Methods and Materials sections of Chapters 2 and 3, we do not believe the duplication is sufficient to combine the Materials and Methods in Chapter 1. Moreover, we believe that combining the Materials and Methods would confuse the reader about which methods were used in each of the two assessments.

Comment: *The study is presented clearly. (Reviewer 3)*

Response: No response is required.

Question 10: Are the findings and the limitations of the analyses correctly stated?

Comment: *No. I have several concerns with the way findings are presented and the conclusions drawn. First, the authors should rethink the ANCOVA approach using % silt/clay. While I agree that substrate size can affect bioassay results and contaminant bioavailability, the reader has no idea how variable % silt/clay really is across the 201 sites used in analyses. Given that the authors limited analyses to poly-eurohaline sites only (a wise move), I would expect particle size to be relatively fine at all sites. If not, couldn't the authors limit analyses to a certain range of % silt/clay and thereby avoid the entire issue? The authors need to present analyses demonstrating the effect of % silt/clay on measurement endpoints to see if it's really a significant factor. What do the results look like if % silt/clay is* *not included in the analysis? After all, the authors don't break out TOC gradients, yet that is at least as important as % silt/clay.*

Second, the Field et al. model relied in part on the data used in this study. The authors need to demonstrate why it is acceptable to use a model calibrated, in part, with these data, as a means to examine chemical-bioassay relationships in this study.

Third, I have the same concerns with the y analysis here as I noted in answer #5 above. Most of the sites are unaffected apparently.

Fourth, the chapter does not present any real data to support the general conclusion on lines 16-19, p. 55. This study found little in the way of organism effects.

Fifth, given the limitations of the analyses mentioned on pp. 59-64, I fail to see how the authors can claim to have seen any relationships of significance. These sediments were likely affected by stressors other than metals (e.g., PAHs in the Elizabeth River) as most of the analyses show. 10-d sediment tests won't pick up less acutely toxic conditions, leaving one to scratch their head as to why a particular metric gives the results it does.

Specific comments follow:

- *Figure 6 - How do you have >100% survival?*
- *Line 12, p. 53 - Are you sure Field et al. used 90% survival to classify toxic sediments?*
- *What is "percent composition metrics", line 5, p. 55?*
- *The regression lines in Figures 7-9 don't jive with the dots in many cases nor do the 95% C.I. How were these regression lines computed?*
- *How is TOC in the sediment a stressor? (line 4, p. 62).*
- *Lines 7-9, p. 62: another explanation is that the 0.5 threshold in the model is wrong.*
- *Line 22, p. 62 - line 2, p. 63: This is stated incorrectly. If SEM/AVS is < 1.0, then metals can not cause toxicity. If the ratio is > 1.0, you don't know whether it's toxic or not.*
- *Lines 21-23, p. 63: You would also expect spurious differences given the large difference in sample size between groups.*
- *Lines 1-2, p. 64 seem to me to be directly antithetical to the statement made in the next sentence. The fact is, the different measures often disagree at many sites. (Reviewer 1)*

Response: *First* - What the reviewer may be forgetting is that estuarine sediments are largely a combination of sand and silt/clay. The % silt/clay content of sediments at the various poly-euryhaline sites ranged from 0.1% to 99.4% and was inversely correlated with % sand (i.e., r = -1.00). The TOC content of sediments at the various poly-euryhaline sited ranges from 0.01% to 7.0%, was correlated with the % silt/clay content (i.e., r=0.77). However, despite this correlation, we were intrigued by the reviewers comment, and went back and added % TOC to the ANCOVA. Depending on the metric, % silt/clay alone, % TOC alone, both variables, or neither variable explained significant variation in individual metrics. We have changed the materials and methods, results, and discussion to reflect these new analyses. Despite this, the community metrics identified for being most sensitive to the sediment contamination changed very little. Therefore, the additional analyses did not change our conclusions.

Second - We have added a sentence to the Methods and Materials that describes why we believe including a comparison that contrasts the classification of sites using sediment chemistry and ambient toxicity tests is appropriate, even if a subset of the paired chemistry and bioassay data used by Field et al. (2002) was from this Virginian Province data set:

As the focus of this research is the relationships between classifications of sites with these two methods, sediment chemistry and ambient toxicity tests, and community metrics, we believe it is appropriate to contrast how these two methods classify the sites.

Third - As discussed previously, the statistic γ is only one measure of the relationships exhibited in the contingency tables. We also discuss a comparison of mean *Ampelisca* survival between the two groups identified by sediment chemistry, and consider how many sites agreed in their classification as affected among those sites identified as affected by at least one individual-based method.

Fourth - Individual effects were assessed based on sediment chemistry or ambient toxicity tests. At 25 of 186 sites, the ambient toxicity tests showed significant mortality, while at 29 of 186 sites, maximum p from the logistic regression models $\geq$0.50. While this results in an unbalanced distribution of sites between the affected and unaffected groups, this is a characteristic of the data set we used. For many reasons, we used secondary data, which is data that was collected for another purpose, in this research. As described in the report, EMAP uses a random-selection approach to identifying sampling sites. As a result, affected sites occur in the data set in proportion to their occurrence across a region. For point source contaminants, such as metals and PAHs, contaminated sites are less likely to be encountered than uncontaminated sites. We will add a discussion of the limitations of secondary data to this section as follows: *While the assessments using toxicity tests and biotic metrics may have been more comparable if the duration of the toxicity tests were chronic,* *this is a limitation of our use of secondary data, which was collected for another purpose. We used EMAP data, and because of decisions made by the EMAP researchers, only data from toxicity tests of acute duration were available. Moreover, the random site-selection approach of EMAP results in sampling of uncontaminated and contaminated sites in proportion to their occurrence across a region. This resulted in the unbalanced distribution of sites between the unaffected and affected groups as identified by sediment chemistry or the ambient bioassays.*

Fifth - In this study, we looked at metals, PAHs, total PCBs, and some pesticides, the contaminants for which logistic regressions were available in Field et al. (2002). Therefore, we are unsure why the reviewer makes his second statement. While total PCBs and pesticides were not analyzed at some sites, we show that the other sites with higher concentrations of total PCBs or pesticides were also contaminated with either metals or PAHs. Despite the limitations discussed, we show significant differences in some community metrics between sites classified as affected and unaffected based on the sediment chemistry or the ambient toxicity tests. Furthermore, a limitation we discuss is that the 10-day tests with *Ampelisca* are acute tests, which may not detect more chronic effects.

Specific comments:

- *Figure 6* - We have modified the sentence in the Section 3.2.4 to make this clear, but we state: “From previous analyses, these test bioassays indicated toxicity if survival was statistically different from ($\alpha = 0.05$) and 80% of survival in the corresponding negative control bioassays (Strobel et al., 1999).” This means that the number of individuals surviving in the test bioassay is compared with the number of individuals surviving in the corresponding negative control bioassay to calculate % survival. If survival in the test bioassay is greater than in the negative control bioassay, % survival will exceed 100%. This is standard practice.
- *Line 12, p. 53* - We rechecked this statement with Sue Norton, a co-author on Field et al. (2002), and the statement about 90% survival is correct. Moreover, she clarified that it was 90% survival in the test bioassay, not relative to the negative control. We have modified the statement to include this extra detail.
- *Line 5, p. 55* - “Percentage composition metric” was meant to be a synonym for the more commonly used “composition metric” (Barbour et al. 1999). We have changed all references to this type of community metric to “composition metric”.
- *Figures 7-9* - We are not sure how this reviewer came to his conclusion that the regression lines and 95% confidence limits do not “jive” with the scatterplots. This seems, at best, to be based on

comparing the lines and the scatterplots by eye. Because of the number of data points plotted on the scatterplots, some data points are hidden by other data points with very similar X and Y values. However, this is a characteristic of most scatterplots with a large number of data points. The reviewer does ask a pertinent question, because we did not explicitly state the software packages used in these analyses. We used PROC FREQ, PROC MEANS, PROC GLM, PROC REG, and PROC UNIVARIATE of SAS (1999) were used in various statistical analyses for this report. We have added specific statements about this as appropriate in Data Handling and Analysis sections of both Chapters 2 and 3.

- *Line 4, p. 62 -* As stated, alteration of TOC in sediments is an effect of excess nutrients in estuarine systems. TOC alone is not a stressor, but its alteration may affect the trophic resources available to benthic invertebrates. We have added "on" to the second and third phrases of this series to make it clear that they refer to "effects".
- *Lines 7-9, p. 62 -* Each of these methods is attempting to achieve the same goal, but each has differing technical limitations that affect its ability to come to the same conclusions.
- *Line 22, p. 62 - line 2, p. 63 -* While we do not state explicitly that an SEMS/AVS ratio > 1 means that the sediment should be toxic because of metals, we see why the reviewer might assume this based on our statement. Therefore, we have altered our statements about the SEMS/AVS ratio as follows:

  The Simultaneously Extracted Metals/Acid Volatile Sulfide (SEM/AVS) ratio exceeded one for sediments from 27 of the 133 sites where AVS data were available. However, this means only that the metals may be bioavailable and not that their concentrations are sufficient to cause toxicity (Hansen et al., 1996). This may be why only four of those sites exhibited toxicity in the ambient toxicity tests, and only three sites had a maximum p > 0.5.
- Lines 21-23, p. 63: The reviewer's statement is incorrect. Moreover, we addressed the concern that with so many individual tests one might expect to see significant differences for a few metrics just by chance in our response to this reviewer's comment on Question 5. This was addressed by use of the sequential Bonferroni technique to correct *p*.
- Lines 1-2, p. 64: The point that this reviewer along with many others in the regulated community cannot seem to get past is that despite the basic relationships among the measurement endpoints used by these methods, other limitations to the methods make them less predictive on a site by site basis. Nevertheless, this does not negate the basic relationship. That is the point of most of our discussion.

Comment: *Yes. I do have a question on the extensive chemical analyses described on Page 41. Would it be of value for this report to include either an http web site or a CD with the chemical methods? Depending on how the report is to be used, such an appendix might enhance the overall utility of the report. (Reviewer 2)*

Response: U.S. EPA has published numerous guidance documents and manuals, such as those cited in the Materials and Methods, that detail analytical methods for chemical analysis of water and sediments. The purpose of the descriptions in the Materials and Methods were to summarize the specific methods used in the two studies where data was obtained for our analyses. An interested reader can find further citations of specific guidance document and electronic versions of many of those documents on U.S. EPA's website. We believe a web site associated with the report or a CD with the chemical methods would not add to and only duplicate what is now available.

Comment: *The limitations of the analysis should be discussed in more detail. Once again, data handling & analysis should be explained in greater detail. (Reviewer 3)*

Response: In response to this reviewer's comments and those of the other reviewers, we have expanded the discussion of data hangling and analysis in the appropriate section in greater detail. Moreover, we devoted much of the discussion to the limitations on our analyses because differences among these methods beyond their measurement endpoints.

Question 11: Does the discussion section of this chapter bring out the most important insights of the analyses?

Comment: *There is not real discussion section per se but the discussion presented brings out most of the important "insights." However, I think the conclusions are generally unsupported by the data and that the regression analyses show little or no relationship between measures. (Reviewer 1)*

Response: As stated elsewhere, and contrary to the reviewer's statement that the regression analyses show little or no relationship between the measures, the statistical analyses do show significant relationships between different community metrics and the classifications of sites based on the methods with organism-level measurement endpoints. Furthermore, these relationships are also shown when continuous variables from the methods with organism-level measurement endpoints are regressed against the community metrics. We extensively discuss the other limitations to the methods make them less predictive on a site by site basis, but this does not negate the basic relationships.

Comment: *Yes. The discussion of the index of correspondence on page 48 (lines 3-6) duplicates what has already been written in Chapter 2. I would delete this section and place into a more comprehensive introduction. (Reviewer 2)*

Response: As discussed previously, we have consolidated most of the introductory material in Chapter 1, and removed it from the introductions to Chapters 2 and 3. However, doing something similar with the Material and Methods for the two studies would be much more difficult. While the use of the index of association, γ, is nearly identical between the two analyses, other statistical methods differ as do field methods, laboratory methods, the use of chemical criteria or guidelines, and the community metrics calculated. While some duplication in the Methods and Materials occurs in Chapters 2 and 3, we do not believe this duplication is sufficient to combine the Materials and Methods in Chapter 1. Moreover, we believe that combining the Materials and Methods would confuse the reader about which methods were used in each of the two studies.

Comment: *The most important points of the analyses are discussed. The limitations (p. 63) are relevant to both case studies. The latter should be incorporated into the Introduction. (Reviewer 3)*

Response: We have added a paragraph in the introduction that discusses in general terms the limitation to our approach:

Several limitations are imposed on our assessment by use of these data sets and by technical aspects of the three methods used for the ecological assessment of contaminant exposure and effects. These data sets were collected for purposes that were different from those for which they are used in this report. As a result, some aspects of their study design are not optimal for our purposes. For example, the ambient toxicity tests conducted in both studies were acute in duration (EPA, 1993; 1994a; 1994b), whereas the results of chronic toxicity tests would have been more comparable to the community metrics, which generally reflect longer-term effects. Also, technical differences among the three methods go beyond the methods' differences in the levels of biological organization used as their measurement endpoints. For example, differences are related to laboratory testing versus field sampling and the selection of test species that are amenable to their use in a laboratory setting. The intent of this report is to address the relationships among the measurement endpoints used by the three methods. However, these aspects of study design and technical differences among the methods are discussed in the following chapters to clarify how they affect the observed relationships among the measurement endpoints.

However, we believe we should leave much of the detailed discussion of the differences among these methods that go beyond their measurement endpoints in the Results and Discussion section of Chapters 2 and 3.

Question 12: Does this chapter line up with the objectives that were stated in the *Introduction* and were those objectives accomplished?

Comment: *Yes, given the caveats noted above and also in answer #8 above. (Reviewer 1)*

Response: No response required, although a response is presented elsewhere to the issues the reviewer brought up under other questions. Moreover, the reviewer's comment in answer #8 refers primarily to a comment in response to Question 6.

Comment: *Yes. (Reviewer 2)*

Response: No response is required.

Comment: *The chapter does "line up" with the objectives and they were accomplished. (Reviewer 3)*

Response: No response is required.

<u>Conclusions</u>

Question 13: Are the conclusions stated in this chapter correct (according to your understanding of the problem outlined in the *Introduction*)?

Comment: *Not really. The first sentence needs caveats - relationships were observed with only a few metrics and over the entire dataset.*
Disagreement among measures at a large proportion of the sites indicates little or no real relationship among measures in these studies.
The discussion concerning metrics and lack of stressor-specificity should acknowledge the recent work done in this regard.
The statement that chronic measures should be more predictive of community-level effects is theoretical at best. This study presents no data either way on this point.
The statement concerning the policy of independent application seems to come out of left field here. Nothing in this entire report really addresses this policy in a direct way. In fact, all of the discussion in the two studies, and prior to this sentence in the Conclusion gives the reader the impression that the community metric results are the final arbiter of effect. All of your analyses were designed to test whether either chemistry or bioassay thresholds agreed with the community results.

The last paragraph of the Conclusions does not follow exactly from everything before it. You went to great lengths to discuss how both ambient bioassays and chemical analyses can misinform one about the true condition. Your study just showed that the biology would be apparently impaired yet bioassays and chemistry tell you otherwise. The strength of evidence analysis idea needs far more discussion in light of the difficulties brought out in these studies.
(Reviewer 1)

Response: We have added a caveat about comparisons at individual sites and about community metrics that are sensitive to the effects of these toxicants.
The reviewer's statement that disagreement among measures at a large proportion of the sites indicates little or no real relationship among measures in these studies concentrates on the comparisons at individual sites. It is this focus on comparisons solely at an individual site basis and not on the overall relationship as revealed by our statistical analyses that has created the misunderstandings about the relationships among these methods.
We have added several citations regarding community metrics and stressor-specificity to the discussion as suggested.
We have added a citation that supports the suggestion that chronic criteria or toxicity tests would be more predictive of community-level effects.

The policy of independent application is discussed in the Introduction as a way that these methods are all used by EPA, and lies at the center of the discussion about the relationships between the measurement endpoints of these three methods. Therefore, consideration of how our conclusions might relate to this policy is appropriate.

As we discuss, the reason that the chemical analyses and ambient bioassay may misinform one about the true condition is that they are much more stressor-specific than community metrics. We are suggesting that one way to use these methods better is to begin with the most general indicator, the community metrics, which may suggest that contaminants are one of several possible stressors. Then the other two methods can be used to verify that the likely stressor is a contaminant and identify the specific contaminant. We have rewritten the first sentence of the paragraph to clarify the connection between this suggestion and the previous conclusions.

These differences in specificity that make these methods complementary might be used in a strength of evidence analysis (U.S. EPA, 2000b).

Comment: *The conclusions follow from the discussions in the two chapters. However, in my opinion, the conclusions do not go far enough. There* *should be a re-statement of the importance of the segmented regression in quantifying logistic data (e.g., those based on thresholds or a criterion or test response, etc). It is very powerful and needs to be re-visited in this section. (Reviewer 2)*

Response: We have added a statement about the use of segmented regression to the conclusions:

Moreover, data sets similar to those analyzed in this study that include both measurements of biological assemblages and of stressors might be used to assess stressor-specific, response relationships and identify thresholds for effects associated with specific stressors. The segmented regression technique used in the analysis of the Colorado REMAP data could be used to identify such thresholds for effects.

Comment: *The conclusions are correct but also should be explained more clearly. (Reviewer 3)*

Response: Based on this comment and the comments of the other reviewers, we have added to the conclusions to explain them more clearly.

Question 14: Are the conclusions correctly derived from the information presented in this document, and does the text of this chapter appropriately refer to those findings and adequately support the conclusions?

Comment: *The conclusions seem fairly disconnected from the findings. There is little or no reference to findings in the Conclusion section. (Reviewer 1)*

Response: We discuss the findings that lead to our conclusions and their relationship to the results observed in the individual studies in the Results and Discussion of Chapters 2 and 3. This conclusion section is meant to summarize these findings and their implications, not repeat the discussion.

Comment: *The conclusions are too weakly stated. There is much more value than stated. One other item that might be included is the fact that such analyses are data intensive and require a large data set.*
The only other component that should be discussed somewhere in this report is the "power of the test," or Type II error rate. How much of a difference among sites or among responses is needed to determine that "they are or are not affected?" A discussion of this would lend itself to the useful suggestion that assessors think about sample size in the conduct of such assessments.
(Reviewer 2)

Response: We have added a statement about our use of larger data sets. We do not think the data sets we have used can be manipulated in any way to address reasonably any questions about the "power of the test" or Type II error rate. Although the EMAP sample design includes revisits, we did not consider the data from revisits to avoid correlations among the revisits. Moreover, sites were revisited from 1 to 3 times and only 12 of 89 sites were revisited in the Colorado REMAP study. In the Virginian estuarine EMAP study, only a few sites were revisited between years. Therefore, we do not have the replication in these data sets to support such an analysis. Moreover, part of the concern over use of these methods at individual sites during routine site assessments is that decisions to assess a site as affected or unaffected are often made based on a single visit or at most a limited number of visits.

Comment: *The conclusions are correct but also should be explained more clearly. (Reviewer 3)*

Response: Based on this comment and the comments of the other reviewers, we have added to our explanation of our conclusions.

Question 15: Are there any other conclusions that can be derived from the findings reported in this document that should be added to those presented?

Comment: *Yes. Need to distinguish between general regional patterns or relationships from those that pertain to any given site. Classification of "impairment" using chemical or bioassay thresholds is an art not a science and the thresholds typically have tremendous uncertainty. If you're going to use biological metrics (which I think is fine) they need to be calibrated for the region and based on some knowledge of responsiveness to stressors of concern. The shotgun approach to metrics used here is not very useful. Other comments presented earlier provide other conclusions that I think should be presented here.*
(Reviewer 1)

Response: While we agree with some this reviewer's above comments, we do not think they appropriately apply to this report. General regional patterns are usually taken into account by classifying sites into regions where those patterns are expected to be similar. The study described in Chapter 2 was conducted in such a region, the Southern Rockies Ecoregion of Colorado. Similarly, the study described in Chapter 3 was conducted in the Virginian Estuarine Province of the Atlantic coast. Therefore, regionalization is not an issue directly dealt with by these studies. The reviewer's second statement about "art" is not appropriate, and we feel is incorrect. As stated in the document, we agree that greater understanding is needed on the responses and sensitivity of various community metrics to different stressors, but wonder how the reviewer proposes that this be achieved if we do not test various metrics individually against different stressors. Moreover, if we are working within a single ecoregion or estuarine province, calibration of community metrics is not needed at this stage, where we are assessing the responses of individual metrics to stressors. Community metrics are more generally calibrated to normalize their range (i.e., make them range from 0-10 or another standard range) when several metrics are combined into an index of biotic integrity.

Comment: *As I stated above, the importance of sample size, the value of segmented regression, and the overall need for establishing a framework with which to implement an assessment is critical and would be of value to emphasize in this report. (Reviewer 2)*

Response: As stated previously, we do not think the data sets we have used can be manipulated in any way to address reasonably any questions about the "power of the test" or Type II error rate. Moreover, part of the concern over use of these methods at individual sites during routine site assessments is that decisions to assess a site as affected or unaffected are often made based on a single visit or at most a limited number of visits. In response to this and other comments by this reviewer, we have added a mention of the use of segmented regression for identification of thresholds with community metrics. As discussed previously, this report is not intended to be guidance, so we are unsure what a framework would be established for.

Comment: *The conclusions are correct but also should be explained more clearly. (Reviewer 3)*

Response: Based on this comment and the comments of the other reviewers, we have added to our explanation of our conclusions.

Question 16: Does this chapter line up with the objectives that were stated in the *Introduction* and were those objectives accomplished?

Comment: *The Conclusion seems to be somewhat removed from the rest of the report. Most of the Conclusion is a rehash of the Introduction in terms of how the various measures differ in their level of biological organization. I'm not sure we're any closer in terms of how those levels relate to each other. (Reviewer 1)*

Response: We believe the conclusions derive directly from the results and discussions of the two studies in Chapters 2 and 3, and go much beyond the Introduction. The intent of this report is to demonstrate statistically these relationships between the measurement endpoints, relationships based theoretically on the heirarchical relationships between the levels of biological organization. We have shown this, although we discuss thoroughly the other differences between these methods that affect our ability to predict effects at individual sites. However, we have added, as is discussed elsewhere, to the conclusions to clarify how they relate back to the previous two chapters.

Comment: *It does not address the third objective (page 5; line 3). The level to which these techniques are predictive or protective has not been established. That the techniques work in mining, freshwater systems is demonstrated. Without a discussion of the power of the test, we do not know the sensitivity of the techniques in further cases. (Reviewer 2)*

Response: We have added the following qualitative conclusion concerning the third objective to the conclusions:

This is why the organism-level effects are only predictive to a limited extent of the community-level effects at individual sites and why these methods frequently differ in their assessment of individual sites.

As discussed elsewhere, we believe questions about the power of the test cannot be addressed with these data sets. Moreover, power of the test is just one factor influencing the comparability of these methods, even for other toxicants. That is why we state in the first sentence of our conclusions that our conclusions apply most specifically to the contaminants addressed in these two studies.

Comment: *The conclusions are correct but also should be explained more clearly. (Reviewer 3)*

Response: Based on this comment and the comments of the other reviewers, we have added to our explanation of our conclusions.

Introduction Revisited (re-read the *Introduction* following your review of the document)

Question 17: Having read the document, would you say its nature, purpose and limitations are accurately described in the *Introduction*?

Comment: *Not exactly. The entire issue of multiple stressors is completely avoided in the Intro yet that is a large source of the difficulties encountered according to the authors. Also, the use or misuse of independent application appears to be unaddressed in the studies. The study is about relationships only, not which measure yields artifacts or true results. (Reviewer 1)*

Response: We have added a discussion of the limitations of our study to the introduction in Section 1.1, Data Sets Used. This discussion mentions the issue of multiple stressors and other limitations imposed by the data sets. By expanding on the Introduction in response to other comments, we have made it clear that the focus of this report is the relationships between the measurement endpoints used by each of these methods that differ in the hierarchical level of biological organization that they measure.

Comment: *Yes, but with the caveat that the Introduction should be expanded to include a larger discussion of the regression and correspondence analyses. These sections could then be cut from Chapters 1 and 2, saving text and space. Generally, however, the introduction does state what is to be analyzed and how, and for what purposes. It does lead us into the report. (Reviewer 2)*

Response: As discussed in response to earlier comments of this reviewer and the other reviewers, we have removed the Introductory material that was originally repeated in both Chapters 2 and 3 and compiled it in Chapter 1. As a result, Chapter 1, Introduction, truly focuses on the objectives of the report. Although we understand that the statistical analyses we used may have some importance and potential application outside this report, those methods were not the focus of this report. This report is not intended to be a guidance document. Instead the results of those analyses are our focus, as our intent was to show that relationships can be shown among the measurement endpoints used by these methods in these two data sets. We believe to move any of the methods used in the individual studies to the Introduction would confuse the reader as to the intent of the report. Moreover, many statistical methods differed in some way between the two studies, and all the field and laboratory methods differed, as the two studies were conducted in freshwater streams and estuaries. To combine the methods into Chapter 1 would also potentially confuse the reader about which methods were used in which study.

Comment: *The nature and purpose are accurately described but the limitations are not accurately described in the Introduction. (Reviewer 3)*

Response: Based on previous comments by this and the other reviewers, we have added a discussion of the limitations of the data sets used in our analyses. This discussion follows:

Several limitations are imposed on our assessment by use of these data sets and by technical aspects of the three methods used for the ecological assessment of contaminant exposure and effects. These data sets were collected for purposes that were different from those for which they are used in this report. As a result, some aspects of their study design are not optimal for our purposes. For example, the ambient toxicity tests *conducted in both studies were acute in duration (EPA, 1993; 1994a; 1994b), whereas the results of chronic toxicity tests would have been more comparable to the community metrics, which generally reflect longer-term effects (Karr and Chu, 1999). Moreover, EMAP generally uses a random-selection approach to identifying sampling sites (Strobel et al., 1999; Herlihy et al., 2000), although both studies included some sites where contamination was known or suspected to occur. While both studies were conducted in regions (i.e.the historical mining region of the Southern Rockies in Colorado and estuaries of the Virginian estuarine province of the eastern United States), where widespread contamination of surface water or sediments is known to occur, the number of sites classified into the unaffected or affected groups was unbalanced (i.e., the number of sites in the unaffected groups was larger than the number in the affected group). Many sites were also potentially affected by other stressors that may not be identifiable by comparisons of chemistry to available criteria or guidelines or by the ambient toxicity tests but may affect community metrics.*

Also, technical differences among the three methods go beyond the methods' differences in the levels of biological organization used as their measurement endpoints. For example, differences are related to laboratory testing versus field sampling and the selection of test species that are amenable to their use in a laboratory setting. The intent of this report is to address the relationships among the measurement endpoints used by the three methods. However, these aspects of study design and technical differences among the methods are discussed in the following chapters to clarify how they affect the observed relationships among the measurement endpoints.

Executive Summary

Question 18: Is the *Executive Summary* easy to read?

Comment: *I don't see an Executive Summary - only an Abstract. It is easy to read. (Reviewer 1)*

Response: We meant this question to refer to the Abstract. The reviewer recognized this, and no response required.

Comment: *There is no Executive Summary in the version I received. Do you mean the Abstract? If so, the abstract is generally adequate, but could be enhanced by mentioning the use of regression techniques. (Reviewer 2)*

Response: Again this question was meant to refer to the Abstract, as the reviewer assumed. We have added to the abstract and these additions to include the use of regression techniques as suggested by the reviewer. An example is provided below:

These same metrics also exhibited relationships with contaminant concentrations in regression analyses.

Comment: *The Executive Summary is easy to read. (Reviewer 3)*

Response: No response is required.

Question 19: Will a reader who does not examine the rest of the document get an accurate view of its contents, key findings, and its limitations from reading the *Executive Summary* alone?

Comment: Not really. The large disagreement among measures at a substantial proportion of the sites is not mentioned. Nor is the degree to which very polluted sites in the dataset drove the relationships observed. The many caveats discussed in the report (e.g., acute bioassays versus chronic chemical thresholds) are not even hinted at here. (Reviewer 1)

Response: We have expanded the abstract to give a more accurate view of the report's contents, key findings, and its limitations as suggested by this reviewer and the other reviewers below. The expanded abstract is provided at the end of the other comments in response to this question, below.

Comment: *No. (Reviewer 2)*

Response: See response to comment by the first reviewer and the expanded Abstract below.

Comment: *In my opinion, a reader will not get an accurate view of the key findings & limitations from reading the Executive Summary (Abstract) alone. Additional and more specific results should be provided. (Reviewer 3)*

Response: See response to comment by the first reviewer and the expanded Abstract below.

In order to use bioassessments to help to diagnose or identify the specific environmental stressors affecting aquatic or marine ecosystems, a better understanding is needed of the relationships among community metrics, ambient chemical criteria or guidelines and ambient toxicity tests. However, these relationships are not necessarily simple, because metrics generally assess measurement endpoints at the community level of biological organization, while ambient criteria or guidelines and ambient toxicity tests assess measurement endpoints at the organism level.

Although a basic hierarchical relationship exists between the levels of biological organization used as measurement endpoints by these methods, quantification of this relationship may be further complicated by the influence of other differences among these methods that affect their sensitivity and specificity to the stressors present at individual sites.

Since 1990, the U.S. Environmental Protection Agency has conducted Environmental Monitoring and Assessment Program (EMAP) surveys of both wadeable stream and estuarine sites. These surveys have collected data on biotic assemblages, physical and chemical habitat characteristics and, in some cases, water and sediment chemistry and toxicity. Among these studies is a survey of wadeable streams in the Southern Rockies ecoregion of Colorado in 1994 and 1995 and a survey of estuaries in the Virginian Province of the eastern United States from 1990 to 1993. Streams in the Southern Rockies ecoregion are affected by contamination from hardrock metal mining, while the estuarine sites may be affected by sediment contamination by polyaromatic hydrocarbons (PAHs) and metals. We characterized streams as metals-affected based on exceedence of hardness-adjusted metals criteria for Cd, Cu, Pb and Zn in surface water; on water column toxicity tests (48-hour Pimephales promelas and Ceriodaphnia dubia survival); on exceedence of sediment threshold effect levels (TELs); or on sediment toxicity tests (7-day Hyalella azteca survival and growth). Estuarine sites were characterized as affected by sediment contamination based on exceedence of sediment guidelines or on sediment toxicity tests (i.e., 10-day Ampelisca abdita survival). The results of these classifications were contrasted by use of contingency tables and a measure of association, γ*. Then, assemblage metrics were compared statistically among affected and unaffected sites to identify metrics sensitive to the contamination. In streams, a number of macroinvertebrate metrics, particularly richness metrics, were less in groups of sites identified as affected by metals with the criteria or ambient toxicity tests, while other metrics were not. Fish metrics were less sensitive to the metal contamination, but this lack of sensitivity is likely because of the low diversity of fish assemblages in these Rocky Mountain streams. Similarly at the estuarine sites, a number of benthic metrics differed between the groups of sites segregated using the organism-level measure, while other metrics did not. These same metrics also exhibited relationships with contaminant concentrations in regression analyses. This variation among metrics depends on the sensitivity of the individual metrics to the stressor gradients of interest as many metrics may not measure the community responses characteristic of a specific stressor. The differences between groups for the more sensitive metrics imply that a relationship exists between the organism-level effects assessed by ambient chemistry or ambient toxicity tests and the community-level effects assessed by community metrics. However, the organism-level* *effects are only predictive to a limited extent of the community-level effects at individual sites.*

Beyond the differences in the levels of biological organization represented by their measurement endpoints, these methods differ in their specificity and sensitivity to different stressors. Criteria or guidelines are specific to the contaminants being measured and assessed and cannot assess contaminants or stressors that are not measured or that lack guidelines for comparison. Ambient toxicity tests should detect effects of any toxicants present and bioavailable, but cannot assess other characteristics of a site that can affect the biotic community. Community metrics are the least specific of the three methods, because they measure directly community-level effects in the native assemblages. Metrics may be selected that are sensitive to a specific stressor, but they also will be sensitive to other stressors, such as alterations in physical habitat that are not addressed by the other methods.

Other factors also affect the relative sensitivity and predictiveness of these different methods. Toxicity tests and chemical criteria or benchmarks based on measurement endpoints that are chronic in duration would be more predictive of community-level effects. Toxicity tests often use one or two standard species, which can be more tolerant of specific contaminants than other indigenous species and would be less predictive of community-level effects than a chemical criterion or benchmark based on a species sensitivity distribution composed of many species.

<u>**General Comments**</u>

Question 20: Please state your overall assessment of the technical quality and scientific accuracy of this document, and provide any suggested changes needed.

Comment: *Please see the many comments above. The technical quality is poor to mediocre in my opinion and the study was not well-designed to address the objectives. Even under the best of circumstances, such field-lab studies are very complex and difficult to decipher. There needs to be a better job selecting appropriate and related measures (e.g., chronic bioassays with chronic chemical thresholds). This is particularly so for the community data: calibrated metrics, known reference site/condition data, and probably multivariate approaches should also be used to address study objectives. (Reviewer 1)*

Response: We have tried to address this reviewer's many comments to the extent possible. However, often this reviewer's comments appear to be completely opposite to those of the other two reviewer's, and his opinion of the methods we tested differ substantially from the other reviewers.

We understand that field-lab studies are very complex and difficult to decipher, and feel the statistical methods we adopted in our approach to this data reflect this. As described in the added material in Chapter 1 on the limitations of our studies, we were limited by the data available in these two data sets, as the data sets were collected for other purposes. Therefore, we did not have the choice of using chronic bioassay data. Other data sets that were reviewed and discarded did not even conduct acute toxicity tests and often did not analyze water or sediments for chemicals that have criteria or guidelines. For the questions we asked, calibrating the community metrics or identifying reference sites or reference conditions would not have contributed to an answer. We believe strongly more testing of various community metrics is needed to understand which metrics are most sensitive to particular stressors to assist using bioassessment data to aid in identifying the stressors causing reduced biotic integrity at individual sites. Moreover, multivariate approaches, although intriguing, are not currently used by U.S. EPA.

Comment: *Overall, I very much like the information presented in this report. I would encourage editing the report so it does not appear as much like two separate papers put together. The methods and analyses sections should be combined into one, where there are similarities. Where there are differences in approach to FW or estuarine situations, these should be dealth with in the appropriate chapter. I would certainly make more of the regression approaches (and limitations of regression). (Reviewer 2)*

Response: Based on this and other similar comments and as discussed previously, we have taken the repetitive introductory material from Chapters 2 and 3 and put it into Chapter 1, which is the Introduction. We believe the methods and analyses differ enough that combining them would be problematic and cause confusion for the reader. We discuss this further in the responses to several other comments.

Comment: *The technical quality and scientific accuracy were fine. The major issue associated with the document is that consideration should be given to a more thorough discussion of certain topics that were explained above. (Reviewer 3)*

Response: We have expanded and clarified the discussion on the topics and sections indicated in other comments by this reviewer. The reviewer has not been very detailed in his comments, so we hope we have identified the appropriate topics that needed more discussion.

APPENDIX A

CHARGE TO PEER REVIEWERS

CHARGE TO PEER REVIEWERS

External Peer Review of the Draft Report, Relationships Among Exceedences of Chemical Criteria or Guidelines, the Results of Ambient Toxicity Tests, and Community Metrics in Aquatic Ecosystems

Background

The report to be externally peer reviewed is a major deliverable of a research program with the goal of improving EPA's understanding of the relationships between ecological assessments conducted at different levels of biological organization and is intended to help decision-makers use ecological assessment results. Three methods are used for the ecological risk assessment of contaminant exposure and effects in surface waters or sediments: (1) chemical criteria or guidelines for the protection of aquatic life, (2) direct toxicity assessments of sediment or water, and (3) bioassessments of biotic assemblages, such as fish, invertebrates, or periphyton.

The objectives of this project were to:

1. Assess the availability of data sets that have used two or all three of the methods to assess sediment or surface water quality at different sites,
2. Compare and contrast statistically the results produced by the different methods at different sites,
3. Assess the extent to which individual-level effects are predictive and protective of the effects at the population level and, in turn, of effects at the assemblage or community level.

This research supports long-term goals for water quality to provide approaches and methods to develop and apply criteria to support designated uses, to demonstrate the application of data, methods, models, and designated use requirements to set criteria for protecting human health, wildlife, and critical habitats from pathogens, toxic chemicals, and habitat alteration in freshwater and coastal systems, and to demonstrate the application of classification schemes, data, sediment-response models, and designated use requirements to set suspended solids and sediment criteria for protecting freshwater stream and coastal systems. The research is also intended to support long-term goals in solid waste and emergency response to develop improved methods, models, and tools that allow managers to more rapidly understand the likelihood of harm that contaminated sites pose to ecological systems and to improve ecological risk assessment tools to support site specific needs.

Charge Questions

This document has been reviewed internally at EPA. This external peer review is the next step in the document review process. As an external peer reviewer, please read the entire document and consider the accuracy of the content, as well as the soundness of the interpretation of the findings presented. Prepare a written response to each of the following seven groups of charge questions/statements (20 charge questions total). Please organize your comments in the same order as the charge questions/statements (for your convenience an electronic copy of this charge has been forwarded to you so you may cut and paste the charge questions/statements directly into your written comments).

Introduction

1. Does the introductory chapter make a coherent statement about the nature, purpose and limitations of this document, and of the research it describes?
2. For those areas within your expertise, is the information accurate, clear and concise?
3. For those areas outside your expertise, is the information clear, concise and easy to follow?
4. In terms of completeness, organization and level of detail, does the information seem to provide an appropriate introduction to the topics covered, for the purposes of this document?

Southern Rocky Mountain Ecoregion R-EMAP Study

5. Is this study presented clearly, and would it be easy to understand for a reader unfamiliar with the three study methods?
6. Are the findings and the limitations of the analyses correctly stated?
7. Does the discussion section of this chapter bring out the most important insights of the analyses?
8. Does this chapter line up with the objectives that were stated in the *Introduction* and were those objectives accomplished?

Virginia Province Estuaries EMAP Study

9. Is this study presented clearly, and would it be easy to understand for a reader unfamiliar with the three study methods?
10. Are the findings and the limitations of the analyses correctly stated?
11. Does the discussion section of this chapter bring out the most important insights of the analyses?
12. Does this chapter line up with the objectives that were stated in the *Introduction* and were those objectives accomplished?

Conclusions

13. Are the conclusions stated in this chapter correct (according to your understanding of the problem outlined in the *Introduction*)?
14. Are the conclusions correctly derived from the information presented in this document, and does the text of this chapter appropriately refer to those findings and adequately support the conclusions?

15. Are there any other conclusions that can be derived from the findings reported in this document that should be added to those presented?
16. Does this chapter line up with the objectives that were stated in the *Introduction* and were those objectives accomplished?

Introduction Revisited (re-read the *Introduction* following your review of the document)

17. Having read the document, would you say its nature, purpose and limitations are accurately described in the *Introduction*?

Executive Summary

18. Is the *Executive Summary* easy to read?
19. Will a reader who does not examine the rest of the document get an accurate view of its contents, key findings, and its limitations from reading the *Executive Summary* alone?

General Comments

20. Please state your overall assessment of the technical quality and scientific accuracy of this document, and provide any suggested changes needed.

If your suggestions include references to published material, please provide a complete citation of the published paper. If any of your comments are limited to particular sections of the document or to particular issues, please be clear what your comments apply to.

DUE DATE FOR WRITTEN COMMENTS: **Friday, February 10, 2006**

When sending review comments via e-mail to ERG, please attach them as WordPerfect 6/7/8 or Word 2000 or later, and save with the appropriate file name extension (.wpd for Word Perfect documents or .doc for Word documents). Send them to Laurie Waite at <Laurie.Waite@erg.com>. If you send your comments electronically, please also express mail or fax a hard copy so we can ensure the electronic data was not corrupted.

If you send a fax for the deadline, please also express mail a hard copy and a copy on diskette to ERG.

Eastern Research Group, Inc. (ERG)
110 Hartwell Avenue
Lexington, MA 02421-3136
Attn: Laurie Waite
E-mail: Laurie.Waite@erg.com

Thank you for your cooperation. Feel free to contact me at 781-674-7362, or Kate Schalk at 781-674-7324 with any questions or concerns.

FORMAT GUIDELINES:

We will submit your comments to EPA exactly as received. Please prepare your comments referring to the above charge questions, organize them in the same order and format them as follows:

TYPE SIZE:	11 point
PAPER SIZE:	8 ½" x 11"
SPACING:	1.5 line spacing
MARGINS:	1" left-hand, right-hand, top, and bottom margins

# Please use a header with your name in the upper right-hand corner of each page of your comments.

# Organize your comments following the order of the charge questions/statements. Use the questions from the charge as the headings to organize your responses. Be sure to provide a response to each question. *To assist you in preparing your comments, ERG has sent you an electronic version of the charge so you may cut and paste each question into your text to be followed by your comments.*

# If commenting on specific information in the document, make sure to **denote the page or section number** the comment refers to.

# Remember to spell out acronyms when first used.

# Avoid incomplete sentences, abbreviations, and terms that might confuse the reader.

# If illustrations or tables are included, be sure that they are suitable for reproduction.

# Submit comments on diskette created in WordPerfect 6/7/8 or Word 97 or 2000 for IBM-compatible computers, or via e-mail with the .wpd or .doc file extension.