Appendix N

Peer Review Document for San Francisco Estuary Institute (SFEI) Report

Author's Response to Peer Review Comments for:

Estimated Sediment Contaminant Concentrations Associated with Biological Impacts at San Diego Bay Clean-up Sites

By Bruce Thompson, Aroon Melwani, and Jennifer Hunt
Aquatic Science Center, Oakland California

July 23, 2009

Review of "Estimated Sediment Contaminant Concentrations Associated with Biological Impacts at San Diego Bay Shipyards"

J. T. Oris, Professor
Department of Zoology
Miami University
Oxford, OH 45056

NOTE: In addition to the below responses, I have provided comments using "track change" tools in MS Word and those comments can be found throughout the document.

General Questions

1. Are the conclusions in the report based upon sound scientific knowledge, methods, and practices?

In my opinion, the report does not really make conclusions, but rather presents a series of options to consider in the negotiation of how to set limits.

It was our intent to present "a series of options". However, some conclusions about our analyses are included (mixtures, relationships with biological indicators, etc).

Based solely on the report, it is not possible to evaluate the soundness of the knowledge, methods, and practices. The report is more of an executive summary than a technical report. A true technical report would have details of the methods and approaches and those are not contained herein. I did not have access to the SWRBC documents. I did obtain and review the publication by Ranasinghe (2008). I am fairly familiar with composite index calculations, but the description of methodology in that paper was not sufficiently transparent for me to know what they did to derive the values.

We did not intend this to be a Technical report, rather a report to the SDRWQCB staff, who are familiar with the background data and SQO methodology. Owing to time and budget limitations, we chose to simply reference the source reports that include details for each study. However, we did include more thorough descriptions of the mean benthic index, and mERMq in the Methods of the final report.

I am not fully immersed in the sediment quality regulatory or policy community and thus I found the prodigious use of acronyms with little attribution to be a bit confusing. Several times, especially with the mERMq index, I could not figure out what it was or what it meant, and it does not seem to be described in the methods of the report.

See response above.

2. Are assumptions of the study valid and reasonable?

There was no specific section or description of the assumptions.

True, there was no specific section. We tried to include language in the report about our assumptions (e.g. bottom of page 1).

3. Are data used in the report reliable and appropriate, and is the treatment of the data defensible?

There was no way to evaluate the reliability of the data since all that is presented in the report is the summary statistics. Again, this is more of an executive summary than a technical report.

See first comment above. Table 1 was added to the final report to show comparability of data.

The treatment of the data is defensible to an extent, but without additional information, data, and technical details, I cannot answer this question.

4. Does the report as a whole support its scientific conclusions and recommendations?

I do not agree with the findings of the correlation analysis and that there did not seem to be a single factor that was common among the Mean Benthic Index (MBI) impacts. I think there was an over reliance on the correlation coefficients and not enough reliance on visual inspection of the data. My conclusion (described in comment windows in the report) is that zinc (or a factor not measured that covaried with zinc) may be an important driver of MBI. Other contaminants of concern included Cd, Chordane, DDT, HPAH, and LPAH. I do not see any relationship between Cu, Hg, Pb, and PCB with MBI in these data displays.

We believe that our analyses needed to be objective and we relied on statistical analyses; visual inspection alone is not adequate. Zinc was indeed strongly correlated and we added a sentence pointing that out (top paragraph, page 6).

Specific Questions

1. Are the statistical methods used appropriately in the study to link chemicals to biological indicators?

Methods for some aspects were described adequately. Most of the methods were references to reports or publications. If one assumes that published methods are appropriate, then they were used appropriately here. Again, since I do not do these analyses on a regular basis, I am not overly familiar with them. In the interest of transparency, I would ask for more details to be presented.

Done, per responses above.

The results of the visual inspection of the single correlation analysis and the PCA/multiple regression analysis do not match well. Even though the models were 'significant' they mostly did not explain an adequate amount of variation in the data. Little attempt was made to compare the results of the correlation analysis with the multivariate analysis. This would have been a good step toward validation of the methods and conclusions.

Correlation were used to determine which contaminants to include in the PCA: All contaminants included in the PCAs were significantly correlated with biological effects (Methods, page 4). Results included discussion of correlations and multiple regression results (Results page 6, Table 6).

2. Is the mean benthic index an appropriate biological indicator?
Given my comments above and my lack of being able to evaluate the methods presented in the report and in the cited publication, I cannot answer this question.

Noted. The final report includes better description of mean benthic index (was Methods), and Figure 2 was added to illustrate the method.

3. The report includes calculated limit values, and for context, other published thresholds. Is that a valid comparison to use, or are there other ways to interpret the limit values.

The table would be better presented graphically and some form of comparative analysis done would help. The data display for this section is confusing and may have errors (e.g. red lines and green lines). I think a modeling approach, using the distributions of data could be used to build a more robust model of limit values.

We did not feel it was necessary to include graphs for all tabulated limits. There are Figures with graphical examples for selected limits, as examples.

Modeling approaches are a good idea, but there could be many ways to do it, and would require discussion and agreement about methods (all with many assumptions). In the end they are all based on the data at hand.

4. Is the ERM quotient used appropriately to develop the empirical limits

associated with sediment impacts?

I cannot figure out how this was calculated or what it tells me based on what is presented in the report. The first mention is in the executive summary and is used without definition or attribution. Is this considered a common knowledge calculation? The next mention is in the Results and Discussion (p. 8) and is used in the context of "mean ERM quotient", but again without attribution as to what it is and how it is calculated.

Explanation included in Methods (top of page 3).

"Estimated Sediment Contaminant Concentrations Associated with Biological Impacts at San Diego Bay Shipyards"

Comments provided by Donald Weston, UC Berkeley

The authors have done a thorough job in doing the analyses they were tasked with doing, but I must admit to having a great deal of difficulty with this report. Some of my difficulties I believe lie in the presentation, which could be clarified, but some are of a more fundamental nature, and go to the heart of why one would even try to do some of the analyses that were attempted or expect success with them. My comments are below, and while I think they are generally responsive to the peer review questions the Region provided, I did not organize my comments in that format.

1. I have few concerns with what is essentially the first half of the analyses, in which a variety of statistical approaches were pursued to identify those contaminants for which the concentrations most closely match the biological responses. My only comment pertains to the geographic scale on which the approach was used. Attempting to identify the primary contaminant structuring the biological response on the scale of San Diego Bay seems doomed to failure. It is far too complex and heterogeneous an environment to expect a single or even a small subset of contaminants to dominate. The approach may be worth trying with just the shipyard data set, but seems more likely to be successful only if applied at a single shipyard. The work done is well described and thorough, but reaches the not-too-surprising conclusion, given the scale on which it was applied, that essentially all the contaminants covary and virtually none stand out or can be ruled out.

Limitations to using the Bay-wide data were clearly stated (Spatial scales, page 3). But we understand that there may be policy reasons for using those limits. Therefore, several scales were included as options for the Board Staff.

2. Given the conclusion of the first half of the analyses, that none of the contaminants can be identified as the driver(s) of the biological response, I don't see why one would attempt to derive compound specific thresholds for clean-up targets or similar purposes. However, that is essentially the thrust of the second half of the report on impact limits. While it is statistically possible to determine, for example, the concentration of DDT above which biological effects consistently occur, if there is no reason to believe DDT is responsible for those effects, then what is the point? It is impossible to distinguish the contaminant responsible from other substances that just happen to covary (as the first half of the paper showed that they all do), so the impact limit is just as likely to be a statistical artifact as it is a biologically meaningful value. It is also quite possible that none of the measured analytes have anything to do with the biological responses, making all the "limits" nonsensical. Those responses may be a consequence of unmeasured substances. I realize the approach used is just a variant of similar approaches going back to the Puget Sound AETs of 20 years ago, but those approaches haven't become any more defensible after 20 years.

We understand that there may be policy reasons for addressing individual contaminants (e.g. identifying a responsible party). We have included statements regarding limitations of interpretation throughout the report.

3. The report acknowledges that it cannot demonstrate causality, but the figure captions of Figures 2A-11A sure sound like causality. "Future samples above the green line *are likely significantly impacted by chlordanes* with 95%, 90%, or 80% statistical probability, whereas future samples below the red line *are not likely significantly impacted by chlordanes* at that statistical level". An additional comment on these figures -- it would be preferable to show the two lines as different patterns, not different colors, for readers that print out the report in black and white.

Captions to even numbered CDF figures beginning with Figure 4 (Page 31) were modified to address this comment. The captions now reflect the proportion of the statistical distribution that may be inferred from the plots, rather than any level of biological impact from the contaminants.

4. I found the half of the report on impact limits to be rather overwhelming. The discussion of the various limits (page 6) is too cursory, but of greater concern is the approach of providing countless variations of the limits. First, there is the distinction between confidence limits and prediction limits, which I still don't really grasp after reading it repeatedly. I understand the brief explanation given (current data vs. future samples), but it is unclear how each is calculated, how they differ conceptually, and when one might want to use one rather than the other. Then for both the confidence limits and prediction limits there are three probability levels. For the prediction limits there are an infinite number of sample sizes possible, though with two sample size choices illustrated. Then there is the choice of whether values are calculated with shipyard data or all San Diego data. Finally, there is the distinction between the highest value of the non-impact stations vs. the lowest value of the impacted stations. The reader is left with a dizzying array of options each with its own corresponding threshold value. The stated purpose of this approach is to provide for flexibility in future decision-making in San Diego Bay. In my view, it is more likely to lead to confusion and minimize any useful guidance. Table 16 provides prediction limits from other studies, and there is a single number for each contaminant in each study. But where other studies derive one number, this report derives, by my count, 27 per contaminant. In my opinion, this work needs to simplify the presentation by picking a probability level, picking a spatial scale, picking confidence or prediction limit, etc. and presenting a much more limited array of choices if it is to be of value to Board staff.

It was our express intent not to make policy decisions on impact limits. Instead this report provides a range of options (albeit 'dizzying') for the Regional Board to consider in the context of their policy decisions. Probability levels, spatial scales, sample size decisions, etc. provide options for clean-up limits depending on the policy decisions made by the Regional board and by discussions with their stakeholders. These options were requested by the San Diego Regional Board.

5. Related to the comment above, the report (page 12) that says if you want limits to apply to all of San Diego Bay, then you should use the Bay-wide data set to derive them. If you want limits for the shipyards, then use the shipyard data. If you want limits for a specific shipyard, then you should use data from that shipyard to derive them. I would think the limit sought should be a maximum concentration that is protective of biota, and thus it is likely to be one number applicable to all of San Diego Bay, not a different number for each part of the Bay. I don't doubt you would derive different limits depending on the data set on which they are based, but I would view those multiple limits only as estimates, with varying degrees of inaccuracy, of what should conceptually be just a single number.

This is a policy decision for the Regional Board to make.

6. The report is easier to follow if one is already familiar with the SQO methodology and terminology. For readers not intimately familiar with the process, some of the explanations are overly brief. An example is the LOE score of ranging from 1-4 and how it becomes an SQO score ranging from 1-5. The presumption of familiarity with SQO jargon may or may not be a problem, depending on the target audience for this report.

Agree. It was difficult to provide a balance of background text and detail needed for the report. However, this report was for the SDRWQCB Staff, who are familiar with the SQO terminology.

7. The sentence, "Especially…" near bottom of page 15 is not a full sentence.

Corrected.

8. Unclear why only half the contaminants listed in Table 16 have empirical thresholds.

The fitted curve (example on Fig. 2) for those not included did not extend above mean benthic index values of 55. Thus, no apparent limit could be determined.

9. Why such odd breakpoints for the mERMq values of Table 17 (for example, 0.128, 0.174)?

This Table was replaced with more rigorous impact limits Tables for three mixture indicators.

Estimated Sediment Contamination Concentrations Associated with Biological Impacts at San Diego Bay Shipyards

Peer Review of ASC Draft Document

Ronald S. Tjeerdema
Department of Environmental Toxicology
University of California, Davis

May 29, 2009

BACKGROUND

Using the Sediment Quality Objectives (SQO) approach, the San Diego Regional Water Quality Control Board recently characterized the sediment quality at a number of sites in San Diego Bay. Since the SQO approach does not correlate contaminants with their potential impacts, the objectives of this study (as listed in the document) were to: 1) identify contaminants of concern that may be adversely impacting the benthic community and causing sediment toxicity, and 2) estimate statistical impact limits for contaminants of concern. Below is a brief critique of the document, focused primarily on the toxicological soundness of the approach and secondarily on the appropriate use/interpretation of statistical analyses.

REVIEW COMMENTS

Overall, the document is very well organized and written, and it generally meets the above goals of the study. It also represents a reasonable descriptive report of the ASC study, the data analysis, and the resultant conclusions. Below are a few specific comments for consideration.

1. Methods Section, Data Subsection:

   While the number of samples and the locations of their origin seem reasonable, they were collected by at least three different programs. A brief summary of the similarities/differences of the sampling techniques is in order to confirm the appropriateness of their comparison in this study.

This report was for the SDRWQCB staff who are familiar with the studies. We reviewed each source report, and deemed that methods were appropriate and decided that comparability was not an issue. Time and budget did not allow for a written comparison. We believe that reference to each study was sufficient.

2. Methods Section, Sediment Chemistry:

   By the nature of this type of study – focusing on a limited number of analytes – correlations to adverse effects will possess a certain degree of uncertainty. This is

inevitable, as no single study can take into consideration all possible candidate contaminants (or their potential synergistic combinations). However, please briefly address this limitation in both this section and later in the Conclusions section (see below).

Also, the quality of analyses performed by the different programs in reporting the sediment chemistry should be briefly addressed, as the data is potentially being weighted equally in the study. For instance, what were the significant similarities/differences in the methods, detection limits, precision, accuracy, etc?

We have described the limitations to interpreting the results in both the Methods and Conclusions. We also added Table 1 that compared chemistry detection limits.

3. Methods Section, Biological Indicators:

Nice recognition of the limitations of using the four category values and the creation of a mean benthic index to attempt to address the limitations.

Noted.

A brief description of the test conditions and parameters used in acquiring the percent survival of *Euhaustorius estuarius* is important to gauge the quality of the data derived. For instance, were the tests performed under static, static renewal or flow-through conditions, were the sediment exposure concentrations analytically verified or nominal, etc? The strength of the conclusions of this study ultimately hinges on the quality of both the sediment chemistry and biological impacts data. Thus, in looking over the results, one of the first questions that arise is "how good is the data that was used?"

We included a reference to the ASTM procedures for the amphipod test, and brief text about the test in the Methods.

4. Methods Section, Interpretation:

The approach for designating "important contaminants" seems reasonable considering the limitations of such a study, as is the designation of a contaminant as a probable cause of a biological impact. Nice use of other published sediment thresholds (Table 16) to put the impact limits presented in this document into perspective.

Noted.

5. Results and Discussion Section, Relationships Between Sediment Chemistry, Benthos and Toxicity:

Please see comments above regarding sediment chemistry and toxicity data quality. In general, the correlations and PCA analyses appear to be appropriate and well described.

See response above.

6. Results and Discussion Section, Impact Limits:

   Calculation of the impact limits as set out in the document (using stations characterized by their SQO scores) seems reasonable, as does the suggestion that they be used as a guide to potential sediment cleanup activities.

   Noted.

7. Conclusions:

   Although the correlations between sediment impacts and contaminants included in this study were significant, they may not actually have a causative relationship. By the nature of this type of study (see comments above), relationships can only be drawn for the limited number of contaminants analyzed for. However, as appropriately stated later in this section, there are potentially many other chemicals present (pesticides, PBDEs, etc.). Please stress more directly that there is a significant possibility that some or all of the correlations derived by this study may be coincidental, and that the real causes may be due to the presence of contaminants not analyzed for. Along with data quality, this represents another significant limitation to this study and should be more directly addressed.

   Regarding data quality, see response for comment 1 above. Chemistry data comparability was summarized in (new) Table 1. We have included cautions about interpretation through out the report.

**Authors note: It is our understanding that this review was to be considered an "informal" review, requested separately by the Regional Board. This reviewer's comments were substantial, and gratefully received. We include our responses here for completeness.*

Technical review and comments on "Estimated Sediment Contaminant Concentrations Associated with Biological Impacts at San Diego Bay Shipyards"

Steve Bay, SCCWRP
June 24, 2009

Summary: This report uses a dataset of sediment quality information from San Diego Bay to evaluate various strategies for relating sediment chemistry to biological response or sediment quality condition. Two objectives were stated: 1) identify contaminants of concern likely to be impacting benthic organisms and 2) estimate statistical impact limits for the contaminants. The methods to identify contaminants of concern used a combination of correlation analysis, principal components analysis (PCA), multiple regression, and comparison to ERM values. Calculation of statistical impact limits were based on confidence limits and prediction intervals at various percentiles. In addition, an empirical threshold for benthic impacts was estimated using a graphical method.

Overall Impression: The statistical methods used in the analysis are basically sound and consistent with standard practice, with the possible exception of the set up of the PCA and regression analyses. The various impact limit values appear to have been calculated correctly based on the methods described, but not enough information is provided to allow even a cursory check of the values. Overall, the information presented should be a useful technical supplement to the development of cleanup levels in the Bay, but the report itself is seriously deficient and flawed in its presentation of the material. I would not suggest that the report be released for public distribution without substantial revision. The critical aspects of the report are summarized below and additional comments have been inserted in the report file (transmitted with this document).

Follow-up conversations with Steve resolved his concerns and we included solutions in our final report. The major changes are detailed below.

Major technical issues need revision
1. The report contains a number of inaccurate statements regarding the characteristics of the study sites (not all shipyards), availability of data, and interpretation of the impact limits. There are numerous instances where information in the text, tables, or figures do not agree (detailed in comments on document). These should all be corrected and the results carefully checked to insure that other errors are not present.

We have addressed all of these concerns: changed shipyards to clean-up areas; clarified data concerns, improved the text on impact limits (some of these are detailed below).

2. The analyses and interpretation to identify the most likely contaminants to be causing impacts is not based on a sound rationale. For example; the PCA analyses indicate that mixtures of contaminants cannot be statistically broken up into individual constituents, so use of the PCA/multiple regression results as a criterion for identification of individual contaminants of importance is inconsistent with the analysis. In addition, use of ERM values as a criterion is unreliable to identify contaminants likely to be most important as the ERMs have not been shown to predictive of causation in California. I do not think that the narrowed list of contaminants of concern is based on a sound interpretation of the data.

Steve recanted this statement about PCA analyses based on a misunderstanding of how the analyses were used. We removed the use of ERM value as a criterion, which had the effect of including all contaminants at one or another of the spatial scales.

3. Additional supporting information for the impact limit calculations should be provided. Tables 6-15 provide numerous upper and lower limit values, but there is not context to relate them to. The mean concentration of contaminant associate with each group/scale should also be provided so that one can evaluate the relative size of the intervals. Furthermore, the composition of the shipyard gradient group is not described. The shipyard gradient group may be the most significant set of data, yet no information is provided on what samples comprise it. Since the Chollas and Paleta sites (especially Paleta) likely have different contaminant sources than the shipyards, I don't understand the rationale to potentially group these two areas with the shipyards.

Means and samples sizes for SQO score < 3 and SQO score > 2 have been provided in Table 7 (Page 19) at the spatial scales evaluated for each contaminant. The mean values may be used with the corresponding impact limits to infer the width of each limit. Further, the sites that comprise the "gradient group" (referred to as site-specific in the report) are shown on Figure 1 (Page 28) and in Table 6 (Page 18). Sites that met the two criteria in Table 6 for a contaminant were included in the site-specific analyses.

4. Better documentation and corrections are needed for Table 16. This table shows values for an empirical threshold calculated in the study, but no description of the methods or presentation of the analysis results is given. Thus, there can be no evaluation as to the validity of these values. In addition, the upper prediction limits attributed to the SCCWRP/SPAWAR study are inaccurate and apparently based on preliminary results. Values from the 2005 final report should be used instead. Note the SCCWRP study calculated metals values based on fines normalization and this factor should not be ignored in the revision.

Table 18 (Page 24) was modified to include the correct values from SCCWRP (2005).