COMMENTS ON
USEPA RESPONSIVENESS SUMMARY
FOR VOLUME 2C-A LOW RESOLUTION SEDIMENT CORING REPORT

Paul Switzer

Responses to my earlier comments were disappointing. Here are general observations, which are illustrated later in this document.

1. Responses frequently assert that assert that "geochemical knowledge" was used to justify unwise statistical procedures, particularly in the design of the 1994 survey. Good statistical practice is needed to draw inferences from sample data and requires carefully designed sampling procedures which are free of bias and purposive selection. Statisticians are distrustful of the "I know better" arguments.

2. Some responses that invoke statistical concepts are inarticulate and meaningless as understood by statisticians, suggesting that responsibility for replying to my earlier questions and criticisms may not have been entrusted to professional statisticians. This is disheartening and reveals a misunderstanding about the central role of statistical inference when drawing conclusions from survey data.

3. Responses often acknowledged the validity of my earlier criticisms but argued that other evidence nevertheless supports the same conclusions. My criticisms were based on what EPA had presented as its argument and it does not seem scientifically balanced to later pick and choose which of EPA's earlier findings should then be de-emphasized.

4. Responses often acknowledged the conceptual validity of my criticisms but then claimed, with *no substantiation*, that they are not of practical importance. My original critique addressed statistical issues that were central to USEPA's conclusions.

The attached comments use the indexing scheme presented in USEPA's Responsiveness Summary.

LG-1.29

This point dealt with the important issue of *which* 1984 sampling sites should be selected for matching in the 1994 sampling survey. I made the point that a statistical design was absent. The response states, *inter alia*, that samples were placed close together because spatial correlation was clearly evident. This is exactly the opposite of what should be done when spatial correlation is present, as any geostatistician will tell you. The response also dwells on the need to do matching which is completely beside the point of my criticism.

LG-1.30

The response states that dredge boundaries were made available after the sampling was completed which reinforces my point of purposive combination of data, as illustrated in the comparison of Figure 4-21 with Figure 4.22.

LG-1.31

Nearly 40% of the collected data were excluded from the analysis, using criteria that could be related to the presence or absence of removal. The response invokes "knowledge of geochemical processes" for the data exclusion criteria that were used.

LG-1.32

I had criticized exclusion of low concentration data and trimming of data to achieve desired distributional shapes. The response emphasizes my original point in saying "The exclusion of low-level samples..is an attempt to exclude samples wherein the expected relationships are unlikely to apply." An astonishing response, indeed.

LG-1.33

Here I was comparing Figure 3-2 with Figure 3-8 to demonstrate how data selection can change weak associations to strong associations. The confusing response states "The analysis was done to confirm an already proven relationship", which is off the point and raises the question of the need to confirm a *proven* relationship.

LG-1.34

The response acknowledges the validity of my criticism of the SSW correction for comparing the early and later surveys, but claims that "there was no other basis to establish the sediment density". Well, does this make it good?

LG-1.35

I criticized the failure to obtain cross-calibration information to assure that the earlier and later surveys were using the comparable yardsticks. The response excuses this lack by stating that "reconstruction of the original techniques is difficult". I'm not sure if this is an answer or an apology.

LG-1.36

I referred to the unaccounted error associated with extrapolating grab samples to 12-inch depth. The response acknowledges the difficulty but states that "it is unlikely that the main conclusion.. will be directly affected". How do we know?

LG-1.37

This point refers to the "regression fallacy" wherein the resampling of high concentration sites is likely to result in lower concentrations, in the absence of removal processes, when we have either measurement error or short-scale spatial variability. The response clearly acknowledges the fallacy but adds a considerable amount of obfuscating material, such as the need to resample hot spots, which has nothing to do with the regression fallacy.

The response also implies that areas sampled in 1994 show less short-scale spatial variability compared with the more comprehensive 1984 survey. The response cites, for example, that in for these restricted areas the spatial correlation at 10-foot separation is 0.47, a number derived from fitting gaussian correlation model. In the first place such a correlation could be regarded as low rather than high. Second, a straightforward empirical correlation should have been reported rather than a model-based correlation. Third, Table A-1 of the Responsiveness Summary indicates substantial local variability ["nugget effect"] for most subreaches. Finally, the response states that the regression fallacy "could result in a slight high bias in the estimated amount of mass loss". Whether the bias is slight or severe is not documented by any specific statistical calculation.

LG-1.38A

I had pointed out the sensitivity of the lognormal hypothesis and its MVUE. The response states that "minor deviations from a true log-normal distribution introduce minor errors". On the contrary, the MVUE is not robust to departures from lognormality. The response also contains meaningless statements such as "the use of the MVUE is justified in light of the greater probability that the underlying population is log-normal" [greater than what?]. Another example is "all of the hot-spots have a rather high probability of a log-normal distribution" which shows a lack of understanding of the meaning of significance probabilities and reverses its intended usage.

LG-1.38B

The error that I pointed out in the estimating equations is acknowledged.

LG-1.38C

While the respondent agrees that ancillary parameters are potentially important, it is argued that such data were not used in the analysis because they were not available for portions of some data sets. If, indeed, ancillary parameters are important, then they will confound the PCB comparisons, i.e. observed PCB differences over time could then be largely due to differences in ancillary parameters over time. The opportunity to investigate confounding possibilities, in situations where the data were indeed available, was ignored, and the confounding possibilities are not considered or discussed.

LG-1.38D

I criticized the reporting of simple correlations without regard for the information provided by potentially confounding variables. The response acknowledges that such a multivariate approach to correlations "would be interesting to complete" but argues that it would be peripheral. Then, why were the simple correlations presented in the first place? Referencing related computations in other publications does not remove the burden of doing professional statistical analyses for these data.

LG-1.38E

I criticized an unsubstantiated use of a factor of 2 to judge significant differences. The response states that this is an empirical observation derived from some statistical comparisons that was extrapolated to other comparisons.

LG-1.38F

The response agrees with my criticism that comparisons based on upper confidence limits can reflect differences not related to mean PCB levels but nevertheless repeats the meaningless comparisons. The only offered justification is the claim that the 1991 Phase I report provides only 95% confidence limits.

LG-1.38G

My comment noted that an assessment of variability components was absent, which the respondent agreed would be probably interesting, but not necessary, in light of their assumption that the total variance is reflected in the standard error estimates. However, an assessment of the regression fallacy bias, for example, could only be understood properly with an appropriate analysis of variability.

LG-1.38H

I criticized the failure to address the sensitivity of the delta-M ratio to the arbitrary addition of the number 2. The response claims that there is no effect on the shape of the distribution but certainly it affects every statistic derived from the distribution including means, probabilities, etc.

LG-1.38I

I noted that geostatistical methods were not used, as they should have been, in the estimation of PCB inventories and their associated uncertainty, and no account was taken of the spatial configuration of the sample cores. The response does not address these points. The fact that only hot spots were considered in 1994 where the spatial correlation is claimed to be stronger only enhances the need for geostatistical methods.

LG-1.38J

I criticized the pervasive absence of statistical measures of uncertainty resulting from the combined effects of sampling, measurement error, and interpolation. The response agrees with this criticism but argues that including such information would make tables too complicated! [The response cites an example where uncertainty information for Table 4-8 is presented with Table 4-7, but this information is not the relevant information for Table 4-8.]

LG-1.39A

The response acknowledges that the results of the statistical analyses of the ^{7}Be data are inconclusive, and merely states that, taken as a whole, the results are consistent with the anticipated trend. No further statistical analysis is offered to support this last statement whose meaning is itself unclear.

LG-1.39B

I pointed to the anomaly where no significant PCB loss was found and, on the other hand, a computation based on a ratio index did find statistical significance. The response acknowledges the anomaly but argues that the latter method is better. However, ratio-based statistics can be much worse due to their poor statistical robustness, for example when small differences are turned into huge ratios.

LG-1.40A

I stated that there was no statistical evidence offered to support the unequivocal conclusion regarding lack of burial. My criticism also noted that the depth of the peak concentration in the low-resolution cores could not be established with sufficient precision to decide the issue of burial. The response, which refers to other conjectures regarding the burial issue, did not address my criticisms involving inferences drawn from the low-resolution cores.

LG-140.B

The response does not dispute that the fraction of sample locations with PCB decreases is not significantly different from 50%. Instead, the response points to the new Appendix A where the same numbers are reconfigured based on grouping of sample locations into clusters. See my comment below on the new Appendix A.

LG-140C

I stated that many far-reaching statements were not supported by statistical arguments. The response was that these far reaching statements were somehow geochemical facts. However, these geochemical facts such as the presumed scouring and redistribution of sediments beg the very questions that the surveys were supposed to address. For example, none of the statistical tests were tests about redistribution, yet this is how the test results were unequivocally interpreted.

Comments on Appendix A

The new Appendix A assigns the original sampling locations to 14 "areas" and tries to make comparisons between 1984 and 1994 for whole areas. For each area, sample means are calculated for the two surveys and the ratio of the two means is reported. Statistical inference is then based on the average of these 14 ratios, treating them as a sample of 14 numbers from some distribution. See Table A-9, for example. The whole exercise is absurd.

1. The definition of areas was done *post-survey* and opens the door to yet another kind of purposive manipulation of data. It seems that this was done to overcome the inconclusive results of the original paired-location sampling design. Such agglomeration of the data means that a large difference at one or a few paired-locations could swamp smaller and insignificant differences at all other paired-locations within the same area, vitiating the whole point of the matched pairs survey design.

2. Using sample data to represent a geographic area requires a proper geostatistical analysis which takes spatial correlation into account, and provides an estimate of how well the sample data represent the area in which they are located. Appendix A gives simple averages of data from sample locations, ignores the clustering of data locations within areas, and provides no measures of estimation error.

3. Ratios are not robust statistics when denominators are subject to appreciable percentage errors, as is the case here.

4. The number of sample locations in each of the areas varies between 2 and 30. No account was taken that some mean estimates have much higher precision than others. The analyses appear to treat the ratios from each of the 14 areas on an equal footing. The area with the largest number of sample locations shows a 1994 mean that is *larger* than the 1984 mean. This area receives the same weight as an area with 2 sample locations in the calculation of confidence limits.

5. Confidence limits are reported in Table 6 and Table 9 but for which statistical parameter? These can only be confidence limits for the mean of some population from which the 14 area ratios are a random sample. What could this population be, and how can we regard these 14 ratios as a random sample?

6. There are mathematical inconsistencies between assuming a lognormal distribution for individual sample location data, a lognormal distribution for an area average, and a normal distribution for ratios, whatever the silly "1.2" rule may say.

7. The use of a statistical test [Shapiro-Wilks] to test whether the 14 numbers were sampled from a normal distribution or a lognormal distribution is a waste of time. They were sampled from neither. Furthermore, this statistical test has very little power when used with 14 numbers, so that larger significance probabilities should not be read as indicative of a good fit.