PEER REVIEW

SUBJECT: "U.S. EPA/Army Corps of Engineers framework for evaluating dredged material for the proposed placement at the Historic Area Remediation Site (HARS)"

REVIEWER: Richard C. Swartz, Ph.D.
2601 NW Parker Avenue
Waldport, Oregon 97394
Tel: (541) 563-3695
e-mail: "swartz@newportnet.com"

DATE: August 2, 1998

This review focuses on the specific set of twelve questions assigned to me by the U.S. EPA.

Framework

1. Is the EPA Region 2/CENAN Framework for evaluating bioaccumulation results scientifically appropriate for determining the suitability of dredged material as Remediation Material? If not, describe deficiencies.

The basic design of the Framework is scientifically appropriate. The list of chemicals addressed in Table 1 includes the important chemicals of concern for NY/NJ Harbor Projects. Arguably, one or two chemicals might be added or deleted from the list, but the list is comprehensive in its inclusion of chemicals from a variety of classes that are locally relevant. The use of the 28-d bioaccumulation test is an appropriate methodological foundation for the Framework. This test is widely used in sediment risk assessments. It has been thoroughly peer-reviewed and established as a standard method. The required testing of two species, the clam Macoma nasuta and the polychaete Nereis virens, allows assessment of interspecific differences in bioaccumulation to be assessed. The basic comparative strategy in which the tissue concentrations observed during the 28-d test on project material are compared against established benchmarks to determine if the project material is acceptable as remediation material is fundamentally sound. However, the scientific defensibility and efficacy of some of the comparative benchmarks are uncertain, as explained below.

The first comparative benchmark in the Framework is tissue concentration observed during 28-d tests with material collected from an established reference site. If the tissue concentration observed in the 28-d test with the project material is less than the reference, the material is deemed acceptable for remediation with respect to a particular chemical. If it is greater than the reference, further evaluation is required. This is an appropriate first level determination.

The second level of comparison involves FDA Action Levels and Regional Matrix/Dioxin values. There are several problems with each of these benchmark comparisons. First, FDA or Matrix or

Dioxin values do not exist for 53 of the 65 chemical parameters of concern listed in Table 1. Second, the seven FDA Action Levels seem extraordinarily high relative to other benchmarks (see answers to questions 6A and 6B below). Third, the four Regional Matrix values are based on an eighteen-year-old technical derivation which is of questionable validity. For example, the Hg Matrix value of 0.2 mg/kg is simply the mean of 16 tissue concentrations in specimens collected somewhere in the New York Bight (including "in and around the dump site") represented in 4 species reported in 6 papers published between 1972 and 1980. There was no standardization of survey or analytical methods among these six investigations. Only two of the six papers were peer-reviewed journal articles. Neither of the two standard 28-d bioaccumulation test species is included among the four species on which the Hg Matrix Value is based. The Hg Matrix value of 0.2 mg/kg is clearly not an effects based benchmark. It is 5x to 6x greater than the background concentration of Hg in clams and polychaetes near the dump site (Table 1, Column 16-17). In my judgment, there is no scientific validity to the use of this Matrix Value for comparison with the results of 28-d tests with Macoma and Nereis. Similar criticisms could be made of the other Matrix Values. The final problem with the Matrix Values is the decision logic evident in Figure 1 of the Framework. If the tissue concentration observed in 28-d tests with project material is less than the Matrix Value, Figure 1 indicates that no further risk evaluation is needed for that chemical. Given the scientific uncertainty about the validity of the Matrix Values, they should not preclude further risk evaluation using other benchmarks.

The third level of comparison is the "Risk Evaluation" as identified in Fig. 1. The Risk Evaluation includes a number of comparative benchmarks. The benchmark that is consistently and substantially lower than other Risk Evaluation benchmarks is the background tissue concentration for both the clam and worm (Table 1, columns 16-17). I describe below in my answer to Question 19 why the background values should not be compared with 28-d test tissue concentrations. Briefly, they are based on resident species that show order of magnitude interspecific variation in tissue concentrations at the same site. Further, the resident species on which the background values are based do not include the standard 28-d test species, so the uncertainty of interspecific extrapolation precludes valid comparison. Fortunately, there is a relatively inexpensive, and scientifically defensible way to establish appropriate background values. The 28-d test with both species should simply be applied to ~ 10 representative sediment samples collected in the background of the dump site (i.e. near but not immediately adjacent to the dump site boundary). This will establish relevant background tissue concentrations that can be unambiguously compared with results of tests with project material. No interspecific comparisons would be required. No laboratory test vs field collection comparisons would be required. No steady state adjustments would be necessary.

How would the results of comparison to 28-d test background levels be used? First, it is important to realize that it is not possible on the basis of existing knowledge to draw "bright lines" that discriminate levels of adverse ecological effects on the basis of bioaccumulation data for most, if not all, of the 65 chemical parameters of Table 1. Arguably, critical body residues can be estimated for a few chemicals, but there is great uncertainty about these estimates. Effects assessments can be based on toxicity and faunal surveys. Bioaccumulation data can be used to ensure that the HARS remediation actually reduces the bioaccumulation of toxic materials from contaminated sediments. Three benchmarks are needed, all based on 28-d tests with both species: 1) the reference benchmark, as currently

incorporated into the Framework, 2) the background benchmark, as described above, and 3) the HARS benchmark, established on the basis of 28-d tests with ~ 10 representative sediment samples collected within the dump site itself. The intention of the remediation will be achieved with respect to bioaccumulation if EPA/COE establish, as a matter of policy, that a project material can be designated as remediation material only if the tissue concentration of every chemical listed in Table 1 as determined in 28-d tests with two species is less than a concentration equal to background plus 25% of the difference between the background and HARS benchmarks. Thus, if the background value for chemical x is 8 mg/kg and the HARS value is 48 mg/kg, the critical value is 8 + 0.25(48-8) = 18 mg/kg. This rationale is subject to the criticism that it is not effects-based, but I submit that an effects-based benchmark for all 65 chemical parameters of Table 1 is impossible. The advantage of this method is that it is understandable and technically defensible from the perspective that it will unequivocally reduce bioaccumulation of toxic materials. The strategy can be coupled with annual monitoring of effects parameters at the reference, background, and HARS sites. If bioaccumulation, toxicity, and biological community effects do not decline over time, the EPA/COE can reduce the critical value to the actual background value or even the reference value.

3. In conducting the integrated effects evaluation using the types of data provided by the applicant, which of the eight factors for LPC compliance listed in the Green Book are appropriate and relevant? How can a quantitative/strategic framework be established to evaluate tissue data for those factors? Considering that comparison to regional Matrix values and site-specific risk values represent case-specific evaluations, is it necessary to conduct the integrated effects evaluation of the bioaccumulation results?

Factors 1 and 5 are of little use since only two species are considered in the 28-d tests. Certainly, exceedence of a standard by two rather than one species is of greater concern, but that quantification is not an appropriate evaluation of "phylogenetic diversity." Factor 2 is of limited utility because the reference comparison is meant to provide a quick evaluation of very clean material. Many chemicals could exceed the reference, but not be a problem if their concentrations are all less than other benchmarks. Factor 3 is difficult to assess because the reference concentrations may be extremely low and the magnitude of exceedence becomes a function of the precision of analytical chemistry. Factor 4 is difficult to assess because all of the chemicals of Table 1 are toxicologically important if their concentration is high enough. Factor 6 (biomagnification) is an important consideration, especially in comparison to a reference or other standard that is not effects-based. Factor 7 is a separate evaluation from the bioaccumulation analysis.

Factor 8 is most relevant to the assessment of LPC compliance. However, as discussed in the answer to questions 1 and 19, the comparison should be to a background based on tissue concentrations observed in 28-d bioaccumulation tests of sediments, rather than tissue concentrations in species living in the vicinity of the disposal site.

I think the integrated assessment boils down to a consideration of the number of chemicals whose concentrations are close to a critical benchmark. If just one of the Table 1 chemicals exceeds a critical benchmark, the project material is not acceptable for disposal, especially as remediation material. The

project material should also be rejected if the concentrations of several (e.g. 5) chemicals are close (e.g. within 10%) of their critical benchmark. The "integration" in such an evaluation is based on the known cumulative effects of mixtures of sediment contaminants. Ideally, it would be desirable to quantify that cumulative effect. Unfortunately, there is no known way to predict cumulative effects of diverse chemicals, e.g. Cd + PCB + dieldrin. The EPA/COE cannot ignore cumulative effects because of the lack of a quantitative model. Hence, a rule such as proposed above (5 or more chemicals within 10% of their critical benchmark) is an appropriate basis for finding that a project material is not acceptable for disposal at the HARS.

Benchmark and Risk Evaluation Values

6A. Are FDA Action Levels useful as upper limit human health benchmarks?

The FDA Action Levels are much greater than all other comparison data in Table 1, columns 14-20 of the Framework. As a practical matter, they would be likely to have little or no impact on the decision process and are therefore of little use as an upper benchmark.

6B. Would the evaluations be improved by omitting comparison of tissue results to FDA Action Levels?

Comparison to the FDA Action Levels is included as part of the Green Book evaluation process and appears to be required by the Dredged Material regulations. Thus, the comparison may be needed as a matter of policy. Although, the FDA Action Levels seem irrelevant to bioaccumulation assessment, they might be used inappropriately to claim that a proposed dredge material is acceptable from a bioaccumulation perspective because it results in tissue concentrations that are only a tiny fraction of the FDA Action Level. Omission of the FDA Action Levels would prevent their misuse in this context.

Calculations

9. Should total PCBs continue to be estimated by doubling the total of 22 congeners or should it be quantified directly using another measure of quantification? What method is most appropriate for sediments in the NY/NJ Harbor area?

I am not an analytical chemist and cannot recommend specific methods for PCB congener analysis. PCB congeners tend to covary in their distribution even though their relative concentrations may change according to source. The 22 PCB congener analytes required in the total PCB quantification include 19 of the 21 congeners recommended in the Green Book for the summation of total PCBs. The list therefore should provide an adequate total PCB quantification that would reflect the distribution of other, unmeasured congeners. In marginal cases, additional analyses should be conducted. The formula for the extrapolation of the sum of the 22 congeners to all congeners (i.e., total PCB = 2.19 x (sum of the 22 congeners) + 2.19) is attributed in Table 4-4B of reference 60 to a 1992 personal communication from T. Wade. That is a very weak source for such an important equation. T. Wade (or someone else) should document the derivation of this equation for the record.

10. Currently, 28-day tissue concentrations of certain organic contaminants are adjusted by some multiplier to estimate the concentration of those compounds had the exposure been of sufficient duration to allow attainment of steady state levels. Are these adjustments appropriate? Should steady state corrections be applied to any other of the listed contaminants? Are there other compounds for which we test that are not expected to approach steady state within the 28-day period?

It is appropriate to adjust 28-d tissue levels to steady state tissue levels before comparison to tissue standards based on chronic exposures. The literature clearly shows that some compounds achieve only a fraction of their steady state concentration during 28-d exposure (Pruell et al. 1993, Lee et al. 1994). Since benthic tissues in the field will achieve steady state contaminant concentrations, correction of the 28-d data is essential.

The correction factors should be derived from 28-d and much longer experiments with the test species used in the standard 28-d test. Thus, the factors for PAHs (McFarland 1995), pesticides (Lee et al. 1994), and PCBs (Pruell et al. 1993) are based on appropriate methods. Boese et al. (1997; ET&C 16:1545-1563 ) reported additional data on PCBs that confirm an average correction factor of about 1 for 13 PCB congeners accumulated by Macoma. I am uncertain about the accuracy of the factor for heptachlor epoxide derived from 32-d tests with fish (Veith et al. 1979), or the factor for 1,4-dichlorobenzene, derived from the de Bruijn et al. (1989) Kow experiments.

The tissue concentrations of dioxins are not corrected for steady state in Table 1 of the Framework. Pruell et al. (1993) demonstrated that Nereis tissue concentrations of 2,3,7,8-TCDD and 2,3,7,8-TCDF were significantly and substantially higher after 120 days than at 28 days of exposure. Steady state correction factors should be derived for those two compounds from the Pruell et al (1993) data. In the absence of better data, the mean correction factor for those two compounds could be applied to other dioxin congeners. Pruell et al (1993) showed that there was no significant difference in Macoma tissue concentrations of dioxins between 28 and 128 days, so no steady state correction is necessary for that species.

11. Is the calculation and use of BaP toxicity equivalence an appropriate way to estimate the potential carcinogenicity of PAHs?

I think it is a reasonable way to estimate carcinogenicity, given the current state-of-the-science. The assumption of additivity inherent in the summation of TEFs reflects current understanding of the effects of PAHs. Toxicological data are scarce and not available for all compounds, but the use of BaP TEFs is probably the best way to estimate cumulative risk. However, the EPA/COE Memo for the Record has ignored the advice of U.S. EPA (1993), the cited source of the PAH TEFs. First, US EPA (1993) says, “These are not proposed as toxicity equivalency factors (TEF)”, but the EPA/COE Memo identifies them as TEFs. This is more than a matter of semantics. Second, US EPA (1993) says, “The list of PAHs is not sufficiently extensive to meet the needs of Programs and Regions.” There is a clear conflict between the uncertainties highlighted in the source document and the proposed application of these numbers.

12. Similar to PCBs, only a subset of those PAHs present in New York Harbor are measured for testing evaluation. How should the remainder be considered?

PAHs tend to covary in their contaminant distributions. Measurement of 16 parent PAH compounds is likely to detect a PAH contamination problem. It is possible that in marginal cases a real problem might be missed if the contribution of other PAHs was necessary to exceed a critical body residue. To minimize this possibility, a couple of substituted PAHs could be added to analyte list, e.g. alkylated phenanthrenes or naphthalenes. Also, a GC/MS scan could be used to detect peaks that might represent other PAHs of concern on a site-specific basis.

General

16. Is use of the Squibb et al. (1991) report appropriate for identifying the contaminants of concern? Are there contaminants which should be added or deleted from the list of contaminants for which we presently test?

Squibb et al. (1991) did an excellent job of summarizing and identifying chemicals of concern for the NY/NJ harbor estuary based on 1990 and earlier reports. There was a substantial literature available to them and I suspect that an evaluation of the more recent literature would not substantially change the list of contaminants of concern. Nonetheless, I recommend that such a literature survey be conducted to ensure that recent studies with more modern analytical methods have not identified additional chemicals that should be added to the list.

There are several chemicals that appear on the Squibb et al. (1991) Table 19 list of toxics of concern for the estuary that are not included in Table 1 of the Memo for the Record. Two chemicals that seem to warrant further consideration are lindane and hexachlorobenzene. Both of these chemicals are included on the list of chemicals of concern for the entire NY Bight. Both occur in the tissues of several fish and invertebrates from the Harbor/Estuary at concentrations that exceed criteria for Category I.B. Pollutants (Squibb et al. (1991) Table 13).

The Squibb et al. (1991) Table 19 also includes a number of methylated naphthalenes, although they are listed as being of concern for sediments only. This class of alkylated PAHs might be added to the Table 1 list to address concerns about the effects of other PAHs (see response to question 16, above).

18. Is test tissue concentration exceeding reference tissue concentration by less than 10X a meaningful evaluative criterion?

No, it is not a meaningful evaluative criterion and should not be used, even as a screen, to assess the acceptability of sediments for ocean placement. There are two principal reasons why the 10X factor should not be used. First, as indicated in the EPA/COE joint memorandum, reference values are variable. If the reference value is very low, the 10X factor may be overly protective, but if the reference value is relatively high, the 10X factor may be underprotective. Second, and more importantly, the derivation of the 10X factor is entirely arbitrary with respect to the potential for

biological effects. Indeed, there probably would be instances where the bioaccumulation from test sediment is < 10X that from reference sediment, but still greater than one or more of the biological standards listed in Table 1, columns 14-20. This is evident in the hypothetical project data of Table 1. The tissue concentration of lead in clams is 1.010 mg/kg for the test sediment, a factor of only 2.5X greater than the concentration for the reference sediment (0.398 mg/kg). According to the 10X screen, lead might not receive further attention. However, the test sediment lead concentration is quite close to the comparison standard for Human Health Non-Cancer risk (1.3 mg/kg, Table 1, column 15). The test sediment lead tissue concentration would exceed that standard if it was as little as 3.5X that of the reference sediment.

I consider the reference material to be more of a procedural control than a standard of comparison. The primary comparison should be between the test sediment tissue concentration and the comparison standards of Table 1, column 14-20.

19. Are the studies from which background tissue concentrations were calculated weighted appropriately? If not, what method is recommended? Is the use of the mean the most appropriate measure of central tendency? If not, what measure should be used? Are the assumptions, presented on page 14 pertaining to comparisons of bioaccumulation in test tissue to tissue concentrations in organisms from the vicinity of the remediation site, valid for evaluating undesirable effects?

The use of data from a single site (McFarland et al. reference) to define background tissue concentrations for clams is inadequate. Also, the McFarland data for four mollusc taxa (Nucula, Yoldia, Mercenaria, "mollusca") is quite variable and strongly influenced by high tissue concentrations in Nucula. For example, phenanthrene concentrations in the four taxa were 8.18, 15.77, 16.20, and 90.51 ppb (arithmetic mean = 32.67 ppb). It seems problematic to compare such data with Macoma used in lab bioaccumulation tests. Clearly, there are order of magnitude differences among mollusc species in bioaccumulation potential that may relate to feeding behavior, substrate relation and other biological factors. If it was reasonable to make such a comparison, the geometric mean would seem to be a better measure of central tendency than the arithmetic mean used to derive the Table 1, column 16 value of 32.7 ppb.

The even weighting of all stations from all studies seems appropriate for the polychaete data. Use of geometric rather than arithmetic means would have minimized effects of extreme values among stations. Data are not presented in reference 98 that allow evaluation of interspecific differences in bioaccumulation among polychaetes for all four studies. The McFarland data sometimes show extreme values among the four polychaete taxa. For example, the Ni concentrations were 0.96, 1.44, 1.50, and 18.07 ppm (arithmetic mean = 5.49 ppm).

The comparison of test results to background tissue concentrations in organisms in the general area of the HARS would make sense if the test species inhabited the area near the HARS or if there was little difference among resident taxa in tissue levels. Unfortunately, the test species are not resident and there are sometimes order of magnitude or greater differences among taxa. Thus it is not true, as claimed on

page 14 of the Memo for the Record, that "When bioaccumulation in organisms exposed to project sediments is not greater than tissue concentrations in organisms from the vicinity of the remediation site (the background levels), this means that placement of the material would not result in bioaccumulation above existing ambient levels in the general area and thus does not have a potential to cause undesirable effects." This statement might be valid for intraspecific comparisons, but it is not valid for interspecific comparisons.

The concept of comparison to background conditions near the HARS is nonetheless appealing. Valid comparisons could be made by collecting a data set for tissue concentrations in test species experimentally exposed to near-HARS sediments following the standard 28-d experimental method.

20. Can baseline tissue concentrations, from appropriate benthic organisms resident to the HARS, be used as standards to determine suitability for Remediation Material as defined above?

No. Specimens resident to the HARS may be exposed to the historic, unacceptable levels of sediment contamination at the HARS. Use of the tissue concentrations in such specimens as standards would tend to perpetuate the historic contamination and defeat the purpose of the remediation.

As explained in the answer to question 1, tissue concentrations determined in 28-d bioaccumulation tests applied to HARS sediment might be used to define a critical tissue concentration above the background level, but substantially less than the HARS level.