DMMP STATUS REPORT

FRESHWATER SEDIMENT GUIDELINES

Prepared by Dr. Teresa Michelsen (Avocet Consulting) and Laura Inouye (Washington Department of Ecology) for the DMMP agencies.

INTRODUCTION

Ecology's current interim freshwater sediment quality guidelines (FW SQGs) were calculated in 2003 using the floating percentile method (FPM), an iterative error rate minimization technique, with paired sediment chemistry and bioassay data available at the time. The data were predominately from western Washington and Portland areas, and there were insufficient data to include chronic toxicity endpoints in the calculations (SAIC and Avocet, 2003).

In 2007, the Regional Sediment Evaluation Team (RSET) agencies (USACE, USFW, NOAA, EPA, Oregon DEQ and Ecology) began the third phase of FW SQG development, which incorporated quality assured data from a much more geographically diverse area, as well as including chronic data. Additionally, DEQ funded an effort to automate the FPM, which will increase its transparency, usability, and portability. Initial draft values are expected in May 2008, with final values in fall of 2008. Peer review and public review will occur through winter of 2008/2009, with adoption as part of the revised Sediment Evaluation Framework (SEF) in mid-2009.

The SLs will be used for evaluation of material being disposed of in freshwater and for the evaluation of new surface material at freshwater sites. This status update and overview of the FPM methodology has been prepared in order to re-introduce the underlying methods used to calculate the FW SQGs so that stakeholders can adequately review and comment on them when they become available.

DATA COLLECTION AND PROCESSING

1. Data Collection

The data set for this effort includes the data originally collected by Ecology in 2002-2003 (see SAIC and Avocet 2002, 2003 for details). Additional data collection was conducted in 2007 to obtain data sets from a broader geographic region (areas of OR, WA, and ID), data sets with chronic bioassays, and more recent data.

2. Initial Data Screening

In assembling the data set, surveys, analytes, and individual data points were screened out if they did not meet certain initial data screening criteria, described below.

Completeness - Surveys and stations were screened out if they had an insufficient analyte list. Although it would be ideal for all stations to have the same analyte list when developing SQGs, that is not possible when using historical data sets. An analyte list consisting of, at minimum, semivolatiles and metals was selected as a general guideline for including a survey or station, consistent with other national SQG development efforts. Metals and semivolatiles both contribute significantly to toxicity in most contaminated sediment data sets, and if these minimum analytes were not available, toxicity would frequently occur in samples without adequate chemistry to explain it. This would lead to an unrealistically high number of false negatives in the reliability analysis, based solely on the lack of a complete analyte list.

Minimum Amount of Data - For development of SQGs, a minimum number of data points is required. A minimum of 30 detected values was chosen as the lower limit for inclusion on the analyte list, based on past experience in development of Apparent Effects Thresholds (AETs) and FPM values. This is the minimum number that is likely to produce a relatively complete data distribution from low to high concentrations.

As there are over 900 data points in the final data set, chemicals with less than 30 detected values are unlikely to be widespread in the region, or to significantly affect the reliability of the SGS. However, those chemicals screened out due to lack of sufficient data will be mapped to determine whether the few detected values cluster at individual facilities, and might therefore be a site-specific concern. In addition, any unusual chemicals present at individual sites or dredging projects should be assessed using freshwater bioassays to ascertain their potential toxicity.

Non-Toxicity – Specific analytes were also screened out for other reasons. Some analytes, such as iron, aluminum, and magnesium, were screened out because they are crustal elements and are naturally present in high concentrations. Certain conventional analytes, such as grain size parameters and acid-volatile sulfides, were screened out because they likewise are not considered contaminants. Other conventional analytes that could affect toxicity, such as ammonia, sulfides, TOC, and grain size were retained for further evaluation.

Chemistry Quality Assurance - Individual chemical data were screened out based on qualifiers assigned during the quality assurance process by the original authors. Data qualified as H, N, Q, X, or R (defined in Table 1 below) were not included in the analysis. Data with these qualifiers were also excluded in

Ecology's previous round of FPM calculations and as part of the Portland Harbor project.

Table 1. Qualifier Definitions for Screened-Out Data

Qualifier	Definition
H	Holding time exceeded (conventionals)
N	Estimate based on presumptive evidence analyte is present in sample
Q	Questionable value
X	Less than 10% recovery
R	Rejected – failure to meet quality assurance guidelines

Bioassay Quality Assurance - Surveys and individual stations with numbers of bioassay replicates below what is considered a minimum standard in modern freshwater protocols (ASTM 2005) will be screened out.

The freshwater ASTM protocols (ASTM 2005) recommend 8 replicates and require a minimum of 4 replicates in order to provide appropriate power under most circumstances. The minimum of 4 is mainly considered appropriate for less rigorous applications, such as trend analysis between years, and is fewer than the DMMP marine bioassay standard of 5 replicates. The data sets remaining in the database after the above screening have at least 5 replicates.

3. Normalization and Summing

To date, evaluations of the reliability of dry weight-normalized SQGs vs. organic carbon-normalized SQGs has shown that the dry weight values have equal or better reliability than the organic carbon-normalized values (PSEP 1988, Ecology 1997). In addition, the use of organic carbon-normalized SQGs leads to implementation difficulties because it is difficult to understand and explain to the regulated community, and because it is inappropriate in some situations with large quantities of anthropogenically-derived organic carbon. Consistent with regional dredging guidelines and all other SQGs calculated after the original marine AETs, it was decided to calculate the SQGs on a dry weight normalized basis.

In the past, SQGs have been calculated both for individual polynuclear aromatic hydrocarbons (PAHs) and for summed dry weight values such as low molecular weight PAHs and high molecular weight PAHs. In recent years, there has been a trend toward using summed values of PAHs in the development of SQGs, as this may better reflect their mode of action and additive toxicity (Swartz et al., 1995; EPA 2000). A PAH workshop was held in June 2007 among the RSET agencies to discuss how best to handle petroleum toxicity in developing SQGs and bioaccumulative guidelines. The participants at this workshop selected the following approach for dealing with historical data sets.

Historical data will be evaluated on the basis of total PAHs, and total petroleum hydrocarbon (TPH) diesel- and residual-range hydrocarbons. This will be accomplished by assembling one data set with the total PAH values, and another data set with the TPH values. These two types of values will be considered alternatives rather than being calculated in the same model run, as PAHs are a subset of TPH. Inclusion of both values in the same model run produces unreliable results for one or both values, as they are not independent of one another. Comparison of the results of these two runs will determine the final approach toward PAHs for the 2008 FW SQGs.

Individual Aroclors were also summed into a single Total PCBs value, and dioxins/furans, total Chlordanes, and Endosulfans were also summed. The following summation rules were used for chemical classes, as selected by the workgroup:

- If all constituents are non-detects, the sum for that chemical class is treated in the same manner as non-detected individual chemicals, and excluded from model calculations.

- If some constituents are detected and others are non-detects, the non-detects are assigned a value of one-half the detected limit and summed with the other constituents.

- Unusually high non-detected values (e.g., due to interference) are not used; instead a value of one-half the standard detection limit for that analysis is used.

- Total PCBs calculated as a sum of Aroclors is an exception to the above summing rules. Aroclors that are undetected are assigned a value of zero. Because Aroclors are already a mixture of PCBs, and individual Aroclor products are frequently used in industrial processes in the absence of other Aroclor products, it cannot be assumed that non-detected Aroclor products are present.

Various methods of dealing with non-detected data were evaluated by the workgroup, including not summing undetected constituents (i.e., setting their value to 0), using half the detection limit, or using statistical methods to estimate the true value. Using half the detection limit was selected for the following reasons:

- This approach is consistent with the required approach outlined in Ecology’s Sediment Management Standards rule and with DEQ’s standard practice. Because regulated parties will be required to calculate their sums in this manner, the SQGs should be calculated the same way so that comparisons are valid.

- It should reduce the variability and the error that would be associated with using zero for non-detected constituents of sums where most of the other constituents are detected.

- It is a simpler calculation procedure than available statistical methods, which would have to be developed, decided upon, and potentially applied differently depending on the distribution of each individual chemical.

- Statistical methods can only operate on distributions (i.e., they replace non-detect values in distributions with non-zero values that cannot be assigned to any individual station). Because the FPM relies on paired chemistry and bioassay values for specific stations, these methods cannot be applied for this purpose.

4. Comparison to Control vs. Reference

Based on the results of SAIC and Avocet (2002), there appears to be no reliability advantage to using a comparison to reference rather than a comparison to control, for this freshwater data set. Freshwater reference areas have not yet been standardized, and the variability of reference stations in the historical data set appears to overwhelm any theoretical advantage they may provide. In addition, many test stations do not have valid reference stations and would have to be excluded from the analysis if comparison to reference is used. Consequently, a comparison to control provides a much larger and more consistent data set to work with in calculating SQGs. Finally, all of the other national SQG sets that have been developed for freshwater have used a comparison to control. Therefore, it was decided to use comparison to control for derivation of SQGs.

This decision does not limit how individual regulatory programs may choose to interpret and use their bioassay data. It is anticipated that freshwater reference areas may be identified concurrently with this report as part of simultaneous RSET efforts, and once this process is completed it may be possible to use a comparison to reference for future updates of the SQGs. However, it is likely that the process may be more difficult than in the marine environment because of the more heterogeneous nature of freshwater environments, and there may not be valid reference areas for all freshwater sites.

5. Bioassay Tests and Endpoints

Eight acute and chronic test endpoints are expected to have sufficient data to calculate SQGs: *Hyalella azteca* 10-day mortality, *Hyalella azteca* 28-day mortality and growth, *Chironomus* 10-day mortality and growth, *Chironomus* 20-day mortality and growth, and Microtox® 15-minute luminescence bioassays (Ecology 2003 protocol).

The first step in performing SQG calculations, once the data have been collected and screened, is the determination of whether adverse biological effects are observed in each sample (called a "hit" if observed and a "no-hit" if not observed). These biological effects levels are also used to interpret the results of bioassay tests conducted to confirm or over-ride the chemical SQGs on an individual project.

The identification of adverse biological effects generally involves a statistical difference from the control or reference plus some threshold of effects, shown in Table 2 below. Quality assurance guidelines for control and reference samples are also shown. In all cases, "statistically significant" means a statistical difference from a control sample at an alpha level of 0.05. Data transformations, selection of null hypotheses, and statistical testing procedures are identical to those currently in use by RSET for marine sediment data (Michelsen and Shaw 1996, Fox et al. 1998).

Table 2. SL1 and SL2 Endpoints for Biological Tests

Test	**QA Control**	**QA Reference**	**SL1**	**SL2**
Hyalella azteca 10-day mortality	C ≤ 20%	R ≤ 25%	T – R > 10%	T – R > 25%
Hyalella azteca 28-day mortality	C ≤ 20%	R ≤ 30%	T – R > 10%	T – R > 25%
Hyalella azteca 28-day growth	CF ≥ 0.15 mg/ind	RF ≥ 0.15 mg/ind	T/R < 0.75	T/R < 0.6
Chironomus tentans 10-day mortality	C ≤ 30%	R ≤ 30%	T – R > 10%	T – R > 25%
Chironomus tentans 10-day growth	CF ≥ 0.48 mg/ind	RF/CF ≥ 0.8	T/R < 0.8	T/R < 0.7
Chironomus tentans 20-day mortality	C ≤ 32%	R ≤ 35%	T – R > 15%	T – R > 25%
Chironomus tentans 20-day growth	CF ≥ 0.48 mg/ind	RF/CF ≥ 0.8	T/R < 0.75	T/R < 0.6
Microtox® decrease in luminescence	CF/CI ≥ 0.72	RF/CF ≥ 0.8	T/R < 0.85	T/R < 0.75

C = Control, CI = Control Initial, CF = Control Final
R = Reference, RF = Reference Final
T = Test Sample

6. ANOVA Analyte Screening

A second screening of the data set has been conducted to remove chemicals that are not apparently associated with toxicity in this data set. This was accomplished by comparing the hit and no-hit distributions for each chemical to determine if they are statistically different using an ANOVA comparison, with various *p* values ≤ 0.1, 0.05, 0.005, and 0.0005 to show increasing degrees of association with toxicity. Experience with the application of the FPM has shown that chemicals with hit and no-hit distributions that are not statistically different do not affect the reliability of the SQGs developed using that data set. This was verified in some early runs on the Portland Harbor project, as well as recent projects conducted for the Washington State Department of Ecology (Ecology) (Avocet 2003), ODEQ (1999), San Francisco Bay, and Los Angeles Harbor.

THE FLOATING PERCENTILE METHOD

1. Overview

In summary, the steps required to calculate SQGs using this approach include:

- Compile and screen synoptic chemistry/bioassay data
- Select toxicity tests and endpoints
- Assign hit/no-hit status for each station/endpoint combination
- Develop chemical distributions
- Select a range of target false negative rates and identify associated optimal percentile values
- Adjust percentiles for individual chemicals upward to reduce false positives
- Identify final SQGs for the data set based on model results and policy choices regarding appropriate false negative rates

The basic concept behind the FPM is to select an optimal percentile of the data set that provides a low false negative rate and then adjust individual chemical concentrations upward until false positive rates are decreased to their lowest possible level while retaining the same low false negative rate. As shown in Figure 1, the y-axis is the percentile of each chemical's overall concentration distribution in the data set, and is not linearly related to toxicity. The green vertical line shows the concentration range for the samples in which toxicity does not occur, and the red vertical line shows the concentration range for the samples in which toxicity does occur. These ranges may overlap due to site-specific or sample-specific variations in bioavailability or toxicity.

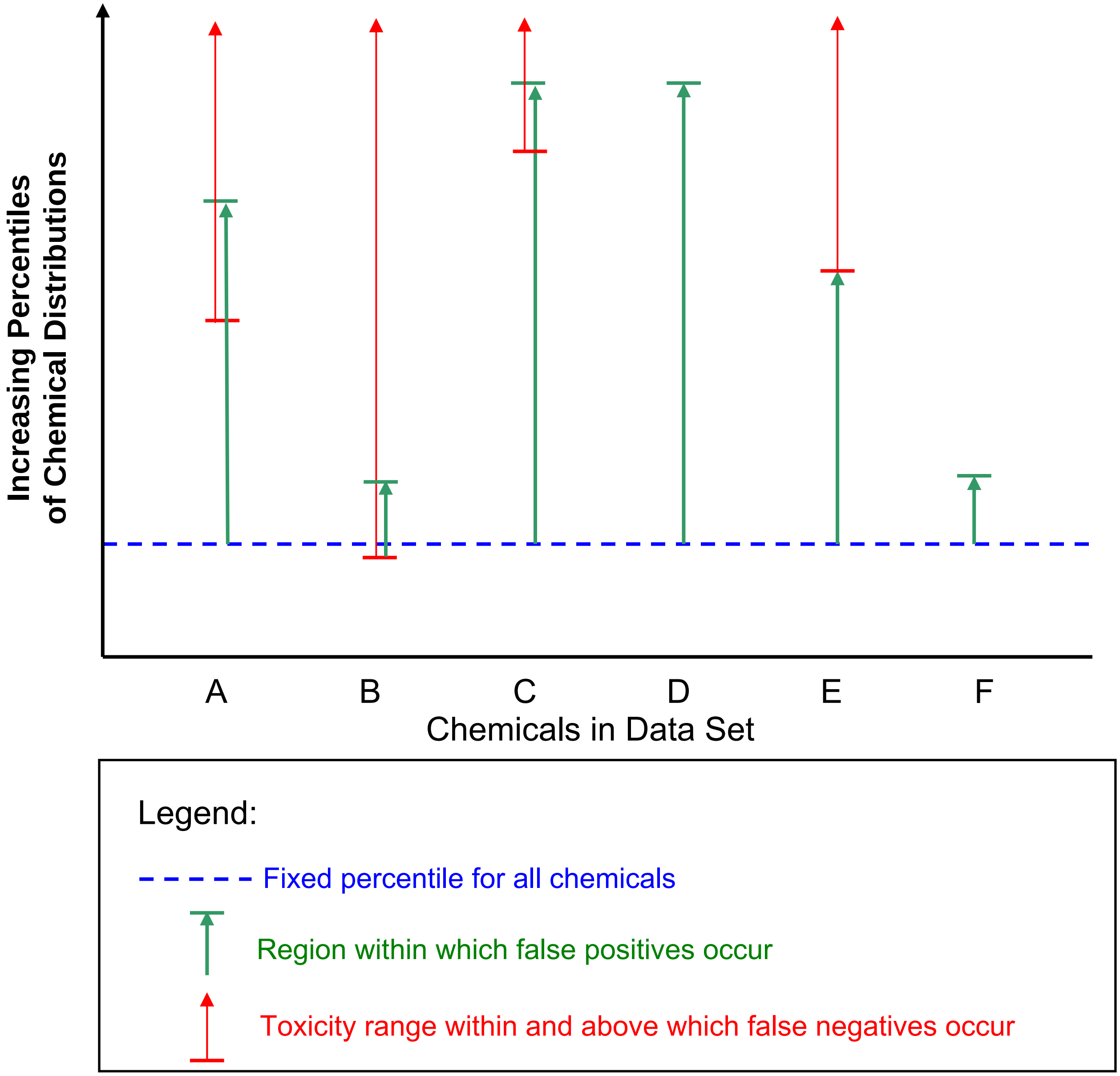
A constant percentile of the distribution that results in a low false negative rate is initially selected for all chemicals, represented by the blue dashed line. The difference between this constant percentile and the lower end of the toxicity range for each chemical is the area between the blue line and the red bar, and this is the source of most of the false positive errors.

The second step is to determine which chemicals are associated with false positive errors in the data set and adjust those concentrations upward until the lower end of their toxicity ranges are reached (red bar). Above this point, false negatives will begin to increase. Above the red bar, both false negatives and false positives may occur, as is shown for Chemicals A, B, and C. This region is the range of concentrations over which sample-specific bioavailability plays an important role in toxicity, and therefore hit and no-hit samples are mixed together, causing both types of errors.

In Figure 1, Chemical B's FW SQG cannot be raised at all because it is already within its toxic concentration range. In any data set, a few chemicals will already be at a toxic level at the initial starting concentration (represented by the blue line), giving rise to a low percentage of false negatives. Some chemicals may show a sharper toxicity threshold (e.g., Chemical E). Others may not appear to be related to toxicity in the data set at all (e.g., Chemicals D and F). These chemical concentrations can be raised to their maximum concentration in the data set without any observed increase in toxicity. However, it may be safer in practice to raise them only to the point where false positives no longer occur (represented by the green bar) or to a similar endpoint such as AETs.

Once each chemical has been individually adjusted upward to the lower end of its toxicity range, the false positives will have been significantly reduced while retaining the same low false negative rate that was initially present. Most chemicals should be at or near their actual toxicity range, rather than at a level arbitrarily assigned by a fixed percentile. In this manner, optimized site-specific SQGs can be developed for a number of different target false negative rates, allowing the trade-offs between false negatives and false positive to be evaluated and a final set of SQGs to be selected.

Figure 1. Floating Percentile Method


Increasing Percentiles
of Chemical Distributions
A
B
C
D
E
F
Chemicals in Data Set
Legend:
Fixed percentile for all chemicals
Region within which false positives occur
Toxicity range within and above which false negatives occur

2. Optimization

Optimization of chemical concentrations occurs through an iterative automated step using Excel macros. The Excel macros use the following approach to conduct the optimization:

1. An appropriate incremental increase for testing is selected for each analyte based on that analyte's complete concentration range (e.g., 1/10 of the difference between the highest and lowest concentration). This increment can now be set by the user.

2. The number of false positives contributed by each individual analyte is calculated, and the chemical contributing the most false positives is selected to begin the optimization procedure.

3. The concentration for that analyte is increased by the chosen increment.

4. After each incremental increase, false negative and false positive rates are recalculated for the entire SQG set.

5. If the false negative rate increases, the chemical concentration is adjusted back down to its previous level and that chemical is "locked in" at that level.

6. If the false positive rate is reduced to zero, the chemical concentration is locked in at that level.

7. If either of the above two conditions is met, or if the number of false positives for that chemical has been reduced below that of another chemical, the macro moves on to the chemical with the current highest number of false positives. If none of these criteria are met, the macro raises the concentration by another increment and repeats steps 4-7.

8. Incremental increases and recalculations continue until every chemical has reached its toxicity threshold or a level at which it has no more false positives.

Through this process, it is possible to identify those analytes having the greatest influence on toxicity in the data set (those whose concentrations cannot be increased without increasing false negatives), and those chemicals having little or no influence on toxicity in the data set (those that can be increased to their highest concentrations with no effect on error rates).

Inspection of the results of the automated process, particularly when various starting percentiles are chosen, also indicates analytes (often metals) with a high covariance in the data set. It may also become apparent that other chemicals, such as PAHs, have relatively little effect individually, but may act in an additive manner to cause toxicity.

The spreadsheets used to develop the SQGs also provide a test area, where candidate SQG sets may be adjusted and finalized, and the results of each change tested with respect to all the reliability parameters. However, recent coding modifications to the automated process have improved the performance of the automated model to the extent that hand-optimization should no longer be required in most cases. This area also allows the operator to enter any criteria set of their choice and test its reliability against the regional data set.

3. Reliability Analysis

Reliability analysis will be conducted following the derivation of the SQGs. The measures of reliability that will be used are listed below:

- **False Negatives:** hits predicted as no-hits/total number of hits
- **False Positives:** no-hits predicted as hits/total number of no-hits
- **Sensitivity:** hits correctly predicted/total number of hits (100% - % false negatives)
- **Efficiency:** no-hits correctly predicted/total number of no-hits (100% - % false positives)
- **Predicted Hit Reliability:** correctly predicted hits/total predicted hits
- **Predicted No-Hit Reliability:** correctly predicted no-hits/total predicted no-hits
- **Overall Reliability:** correct predictions/total stations

False positives and false negatives are the primary measure of predictive errors in the reliability assessment. Each of the other reliability values is related to them in some way. While the performance of any given data set cannot be determined in advance, the workgroup agreed on a set of reliability goals that would guide the selection of the final SQGs, shown in Table 3. Based on the existing interim values in the SEF, the most difficult of these goals to meet is likely the predicted no-hit reliability at the SL1 level.

Table 3. Reliability Goals for Proposed Freshwater SQGs

	SL1[a] (%)		**SL2[b] (%)**	
	SEF[c]	**Goal**	**SEF[c]**	**Goal**
Sensitivity	84	80 – 90	85	75 - 85
Efficiency	75	70 – 80	75	75 - 85
Predicted Hit Reliability	88	70 – 80	77	75 - 85
Predicted No-Hit Reliability	67	80 – 90	84	75 - 85

[a] SL1 = DMMP Screening Level (SL)
[b] SL2 = DMMP Maximum Level (ML)
[c] Actual value achieved for interim SEF freshwater SQGs

To best allow the workgroup and the public to understand the various reliability measures, each parameter will not only be represented as a percentage, but also with the numbers of stations in the numerator and the denominator shown. In addition, figures will be prepared graphically depicting the correct assignments and errors associated with each bioassay endpoint and effects level.

4. Sensitivity Analysis

The relative importance of each individual analyte will be assessed by dropping out that analyte and noting any changes to reliability of the SQG set. This allows an evaluation of which analytes are critical to include in the SQG set, which are of lesser importance, and which may not be needed at all. For those that would not be included in the standard RSET analyte list, a range of concentrations evaluated as part of this effort will be tabulated and provided to project managers to provide a means of evaluating these chemicals at individual sites and projects. Based on this analysis, concentrations within the range evaluated are not likely to be associated with toxicity.

5. Interim Data Runs

The structure of the first set of data runs was briefly discussed within the Working Group, in terms of how many different runs would be needed. The following were identified:

- Total PAHs vs. TPH
- East of the Cascades vs. west of the Cascades
- Removing mining-influenced data if it seems anomalous

MODEL VALIDATION

Peer review of the draft 2003 FW SQGs indicated that validation of the FW SQGs was and remains desirable. At this time, no funding has been identified for the validation studies, but drawing on the previous discussion, a general plan can be developed. Evaluation of the performance of the model in relatively clean, contaminated and intermediate areas was considered a useful concept. Representative areas for each of these three conditions could be selected based on model results, then resampled to determine whether the reliability in the new data set for these areas is similar to that of the model. Similarly, based on model results, the new sampling could be conducted for a subset of the bioassay endpoints and compared to the reliability of those same endpoints in the model. The number of samples that would be required would be determined by the expected reliability of the model once the runs are completed.

ACKNOWLEDGMENTS

This summary was excerpted from a draft report prepared by Dr. Teresa Michelsen, under contract to the Washington Department of Ecology and E&E, on behalf of the RSET agencies. Dr. Michelsen's work has been guided by the Freshwater Sediment Quality Guidelines Workgroup, whose members are as follows:

Keith Johnson, DEQ, Chair
Mike Anderson, formerly DEQ
Robert Anderson, NOAA
Jeremy Buck, USF&W
Taku Fuji, Kennedy Jenks
Lyndal Johnson, NOAA
Laura Inouye, Ecology
Teresa Michelsen, Avocet
Mike Poulsen, DEQ
Paul Seidel, DEQ
Burt Shephard, EPA
Mark Siipola, Portland COE
Dave Sternberg, Ecology (Project Manager)

Invaluable contract assistance was provided by Bill Richards, Blythe Mackey, and Erin Lynch of Ecology & Environment, and James J. McAteer Jr. of QA/QC Solutions. Mike Anderson performed the coding of the FPM method in addition to participating in the workgroup.

Also instrumental in guiding and supporting this process was the RSET management team, including Jim Reese (NW Region Corps), Stephanie Stirling (Seattle District Corps), Dave Sternberg, Dave Bradley, and Chance Asher (WA Dept. of Ecology), and Cathy Tortorici (NOAA). Funding was provided by the Oregon Department of Environmental Quality, Washington Department of Ecology, and NW Region Corps of Engineers.

REFERENCES

ASTM. 2005. Standard Test Method for Measuring the Toxicity of Sediment-Associated Contaminants with Freshwater Invertebrates. ASTM E1706-05. American Society for Testing and Materials, West Conshohocken, PA.

Ecology. 1997. Creation and Analysis of Freshwater Sediment Quality Values in Washington State. Washington Department of Ecology, Environmental Investigations and Laboratory Services Program, Olympia WA.

EPA. 2000. Equilibrium Partitioning Sediment Guidelines (ESGs) for the Protection of Benthic Organisms: PAH Mixtures. U.S. Environmental Protection Agency, Office of Science and Technology and Office of Research and Development.

Fox, D.F., D.A. Gustafson, and T.C. Shaw. 1998. Biostat Software for the Analysis of DMMP/SMS Bioassay Data. DMMP Clarification Paper, SMS Technical Information Memorandum. Seattle District Corps of Engineers, Seattle, WA.

Michelsen, T.C. and T.C. Shaw. 1996. Statistical Evaluation of Bioassay Results. PSDDA Clarification Paper, SMS Technical Information Memorandum. Washington Department of Ecology, Olympia, WA, and Seattle District Corps of Engineers, Seattle, WA.

PSEP. 1988. 1988 Update and Evaluation of Puget Sound AET. U.S. Environmental Protection Agency, Puget Sound Estuary Program, Seattle, WA.

SAIC and Avocet. 2002. Development of Freshwater Sediment Quality Values in Washington State, Phase I Final Report. Prepared by SAIC, Bothell, WA and Avocet Consulting, Kenmore, WA for the Washington Department of Ecology, Olympia, WA.

SAIC and Avocet. 2003. Development of Freshwater Sediment Quality Values in Washington State, Phase II Final Report. Prepared by SAIC, Bothell, WA and Avocet Consulting, Kenmore, WA for the Washington Department of Ecology, Olympia, WA. http://www.ecy.wa.gov/pubs/0309088.pdf

Swartz, RC, DW Schults, RJ Ozretich, JO Lamberson, FA Cole, TH DeWitt, MS Redmond, and SP Ferraro. 1995. ∑PAH: A model to predict the toxicity of polynuclear aromatic hydrocarbon mixtures in field-collected sediments. *Environmental Toxicology and Chemistry* 14(11):1977-1987