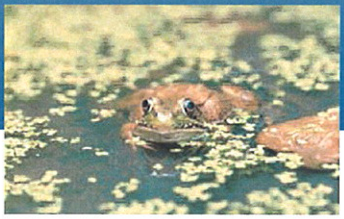
Final Report



An Evaluation of Existing Sediment Screening Guidelines for Wetland Creation/Beneficial Reuse of Dredged Material in the San Francisco Bay Area Along with a Proposed Approach for Alternative Guideline Development

Funded by the California State Coastal Conservancy in cooperation with the Port of Oakland

Prepared by The Germano & Associates Team:

Germano & Associates, Inc.

ExA Data & Mapping Services

TerraStat Consulting Group

CoastalVision Informing Your Decisions

Avocet Consulting

February 2004

An Evaluation of Existing Sediment Screening Guidelines for Wetland Creation/Beneficial Reuse of Dredged Material in the San Francisco Bay Area Along with a Proposed Approach for Alternative Guideline Development

FINAL REPORT

FEBRUARY, 2004

Submitted by:
Germano & Associates, Inc.
12100 SE 46th Place
Bellevue, WA 98006

Working together with:
Exa Data & Mapping Services
TerraStat Consulting Group
Coastal Vision, LLC
Avocet Consulting

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The San Francisco Regional Water Quality Control Board (RWQCB) recently updated the 1992 sediment screening guidelines (SFB-RWQCB, 2000) in support of the Long Term Management Strategy (LTMS) goal that targets 40% of the sediment dredged from San Francisco Bay for beneficial reuse (USACE, 1998). Many of the guidelines, however, defaulted either to ambient sediment chemical concentrations or to national sediment screening guidelines. No information has been available on the predictive reliability of these values when applied to Bay Area dredging projects. The first goal of this project was to measure how well the current RWQCB sediment chemistry guidelines, as well as existing regional and national guidelines, predict actual biological acute toxicity in sediments collected from the Bay. While there are other important lines of evidence to be evaluated when considering the suitability of sediments for wetland restoration projects (bioaccumulation potential, leachate characteristics, geotechnical properties based on specific wetland engineering design), because of the type and quantity of regional data that were available, acute toxicity was the only endpoint that we subjected to a detailed evaluation. Following this evaluation, we employed several quantitative methods to improve the predictive performance of the values, resulting in a new set of suggested optimized guidelines that would be useful for screening sediments for wetland creation/restoration projects.

Because of the historical loss of wetlands in the Bay, wetland restoration is an ideal alternative for dredged material beneficial reuse. The RWQCB guidelines provide two different sets of chemical values for use in screening dredged material designated for both wetland surface and underlying foundation sediments. Because surface material would be in direct contact with wetland flora and fauna, the sediment guidelines for dredged material earmarked for wetland surface placement should be protective of sensitive biological receptors. The guidelines to be used for foundation should focus primarily on preventing potential contaminant mobility to either groundwater or biological receptors.

The first step of the project was to assemble existing regional paired samples with both analytical chemical and biological effects testing. The resulting Sediment Screening Guideline Database (SSGD) that accompanies this report includes tools to recalculate the guideline performance evaluations if new data are added or if different statistical threshold values are chosen by state or federal regulatory agencies. The database includes dredged material characterization as well as regional monitoring data. Both sediment and elutriate bioassay tests are in the database, although only sediment tests were used for guideline analyses. Each data set was evaluated to ensure that it was appropriate to be included for sediment guideline analysis. Final guidelines were developed using results from 337 amphipod (*Ampelisca abdita, Eohaustorius estuarius*) bioassays, the most common metric used for national guideline development, and the majority of data available in the SSGD.

Prior to the performance evaluation of existing sediment screening guidelines (SSGs), the toxicity status of each bioassay sample was standardized using a Reference Envelope approach. Our approach was based on that used previously in the Bay (Hunt et al., 1998). A tolerance limit was calculated for *A. abdita* and *E. estuarius* from valid reference locations used in the Hunt et al. (1998) study. Any mean survival value that fell below the tolerance limit was unlikely to have come from the reference distribution, and was therefore classified as "toxic" in the database, while a mean value that exceeded the tolerance limit was considered not statistically different from reference or "non-toxic."

For the *A. abdita* reference distribution, an acute toxicity threshold based on the 10th percentile corresponded to a normalized survival value of 76.4% as the tolerance limit (e.g., any sample with survival of <76.4% was classified as toxic). For *E. estuarius*, we used an acute toxicity threshold based on the 20th percentile tolerance limit, because a sensitivity analysis indicated that this higher percentile provided a better separation between the contaminated and uncontaminated stations. The final tolerance limit applied to the *E. estuarius* samples corresponded to a normalized survival value of 70.6%.

The synoptic sediment chemistry and bioassay acute toxicity classifications were used to evaluate the performance of the RWQCB surface and foundation chemical values, as well as other national sets of SSGs. The SSG sets included in this effort were selected because of their common use in national or other regional programs (Appendix A). For the majority of the SSGs, a sample was predicted to be toxic if one or more individual chemical guidelines were exceeded, and non-toxic if no chemical guidelines were exceeded. One guideline was not a chemical-specific guideline, but a mean quotient (called the sediment quality guideline quotient, or SQG-Q1) calculated from nine chemicals (Cd, Cu, Ag, Pb, Zn, chlordane, dieldrin, total PAHs, and total PCBs) that, in combination, best predicted amphipod toxicity in national databases (Fairey et al., 2001); for the SQG quotient approach (SQG-Q1), a sample was predicted to be toxic if the mean SQG quotient was greater than 0.10 for surface sediments or 0.50 for foundation sediments.

Because wetland surface material will have direct exposure to organisms, ideal sediment screening guidelines for surface material would avoid predicting contaminated sediments as non-toxic (low false negatives), and correctly predict as many clean sediments as possible (high non-toxic efficiency). In contrast, the goal for foundation material was that the incorrect allocation of clean sediments suitable for surface material should be minimized; therefore ideal SSGs for foundation material would avoid predicting clean sediments as toxic (low false positives), and correctly predict as many contaminated sediments as possible (high toxic efficiency and high sensitivity).

Results of the performance analyses for surface material showed that optimal target rates (expressed as a percentage) were nearly achieved by most of the national guideline sets; however, the number of samples predicted to be suitable for surface material was so low as to restrict the existing guidelines' practical utility. The SFB-RWQCB (2000) surface guidelines had the best performance of the entire set of existing surface SSGs; however, only 57 samples out of 337 were predicted as suitable for surface material. In practical terms, although these guidelines would serve well for safely identifying non-toxic samples for surface use in wetland restoration, the high false positive rate (76%) indicated that a permit applicant should always opt for bioeffects testing instead of accepting the prediction of acute toxicity based on sediment chemistry.

The existing SSGs for foundation material did not perform as well as the surface guidelines. The SFB-RWQCB (2000) foundation guideline had poor reliability results with 18% false positives, 45% toxic efficiency, and only 21% sensitivity. A total of 65 samples (with 45% accuracy) were predicted as suitable for foundation material using these guidelines. The foundation SQG-Q1 quotient predictions performed the best from a rate perspective with 3% false positives and 78% toxic efficiency. However, sensitivity was a low 15%, and only 23 out of 337 samples were predicted as suitable for foundation material. Most of the toxic samples in the database had concentrations below these guidelines, which means that the observed acute toxicity was not likely due to the compounds on the SQG-Q1 list at those threshold levels.

The performance of all of the existing SSG sets suggested that amphipod acute toxic responses cannot be explained by the action of individual chemicals alone; none of the existing guidelines were good predictors of amphipod acute toxicity, and none of them were able to simultaneously achieve a low false positive rate and high sensitivity. We identified several possible reasons for this discrepancy: chemical concentrations do not extend into the consistently toxic range; acute toxicity could be caused primarily by chemicals that are not being measured; acute toxicity could be caused by synergistic or antagonistic effects of measured chemical concentrations; or the measured chemicals are indeed responsible for adverse biological effects, but the test organisms used are imprecise and unreliable indicators of acute toxicity.

Because of the substantial overlap in concentration distributions between the toxic and non-toxic populations, we investigated methods to optimize the existing SSGs. First, we adapted the Receiver Operating Characteristic (ROC) approach for assessment of SSGs (Shine et al., 2003 a,b), both to find chemicals that most accurately predict acute toxicity (acute toxicity drivers), and to attempt to optimize the value of specific SSGs. The chemicals that were most consistently associated with acute toxicity were polycyclic aromatic hydrocarbons (PAHs), but in general the association was, at best, fair. The relatively flat shape of all of the ROC curves indicated that there was no break point between maximizing sensitivity (true positives) and avoiding false positives. Although the ROC approach appeared to have the potential to be a useful method both to quantify and graphically illustrate the relative sensitivity in a data set, the results confirmed that amphipod acute toxicity in Bay samples is not strongly associated with the presence of any individual chemical.

Second, a site-specific method was adopted to improve the reliability of the existing surface and foundation guidelines despite the substantial overlap in concentration ranges. Because of this overlap of chemical concentrations between the toxic and non-toxic populations, we would expect the data to show a costly trade-off between false negatives and false positives. Therefore, we approached site-specific SSG development by using the Floating Percentile method that can simultaneously optimize the false negative and false positive error rates. The Floating Percentile method provided the best combination of performance metrics that are possible for these data.

To establish the optimum guideline values, first a fixed percentile for the data was selected that provided a low false negative rate. Then, individual chemical values were adjusted upward until false positive rates were optimized while retaining the same level of false negatives. The Floating Percentile surface SSGs (Table ES-1) provided a significant improvement in false negatives and non-toxic efficiency. Modest improvements were found for sensitivity, toxic efficiency, and overall reliability. A significant improvement was also obtained for the foundation SSGs for the metrics of false positives and toxic efficiency, with modest improvements in the other performance metrics. The Floating Percentile guideline values (Table ES-1) include the molar sums of PAHs to reflect their additive toxicity based on a narcosis toxicity model. The narcosis model is a general model of toxicity to aquatic receptors based on the presence of organic chemicals in tissues and their disruption of basic cellular functions (Connell and Markwell, 1992; Veith et al., 1983). This mode of toxicity occurs in all species and is dependent only on the total molar concentration of chemicals partitioned into lipid tissues.

Table ES-1. Final Optimized Target Analyte List for Floating Percentile SSGs

Chemical Name	Surface	Foundation
Metals (ppm, dry weight [DW])		
Arsenic	40.0	40.0
Cadmium	0.250	0.620
Chromium	119	320
Copper	50.0	150
Lead	200	200
Mercury	1.18	1.18
Nickel	230	230
Silver	0.280	2.00
Zinc	1,200	1,200
Total molar PAHs (μmol/kg, DW)	6.3	32
Chlorinated organic compounds (ppb, DW)		
Hexachlorobenzene	60	60
Pesticides and PCBs (ppb, DW)		
Total DDTs	250	250
Chlordane	69.2	69.2
Total BHCs	2.0	2.0
PCBs	600	600

An important caveat to Table ES-1 is that these suggested guidelines are based on acute toxicity results and do not account for bioaccumulation potential or chronic effects; regional guidelines for dealing with bioaccumulative compounds of concern have not yet been developed and remain an important priority for San Francisco Bay regulatory agencies.

Many of the Floating Percentile SSGs are identical for both surface and foundation. This occurs for the compounds for which the toxicity threshold is fairly well-defined in the data set. It is important to remember that all toxicity thresholds are conditional on the concentrations for the rest of the chemicals in the mixture. The Floating Percentile process is a multivariate optimization routine that focuses first on the chemicals that are responsible for the most number of false positives, i.e., non-toxic samples that exceed the initial Floating Percentile values. A higher Floating Percentile value is selected only when it results in a lower false positive rate and a constant false negative rate, based on an evaluation of the complete chemical mixture relative to the full set of Floating Percentile guidelines set to date. For example, the Floating Percentile guideline for arsenic goes right to a value of 40 ppm (the maximum concentration in non-toxic samples) for both surface and foundation guidelines, because below this value, false negatives are not affected due to the other chemicals in the mixture exceeding their guidelines, and above this value, false negatives would increase. The toxicity threshold for arsenic in this dataset is well-defined, because it is constant across the range of false negative error rates.

The guideline values that have different surface and foundation numbers were for the chemicals associated with a large number of false positives in this data set. For example, the initial Floating Percentile surface value for cadmium was associated with a large number of false positives, so the surface guideline for cadmium was raised as long as the false negative rate remained unchanged and the false positive rate decreased. If the guideline continues to increase for consecutively higher false negative rates, then the toxicity threshold is fuzzy: a higher guideline would increase the false negative rate, and a lower guideline would increase the false positive rate. Unlike the situation for arsenic, the non-toxic samples with concentrations below the guideline for cadmium were not predicted by other guidelines.

The resulting suggested list of Floating Percentile SSGs included fewer chemicals than the historical RWQCB published values. Although we recommend that permit applicants should still screen sediments for the same suite of contaminants normally required, the results from the all of the analyses are clear: the contaminants in this data set that can be reliably used for screening guidelines are the fifteen chemicals listed in the summary table above. Another important difference between the suggested foundation material SSGs in Table ES-1 and those currently in use (SFB-RWQCB, 2000) is that instead of these numbers being recommended as upper limits, they are instead minimum thresholds that would qualify a sediment to "cross over the line" as foundation material. Figure ES-1 illustrates the difference between the existing foundation guidelines (SFB-RWQCB, 2000) and the suggested values in Table ES-1; if these suggested guidelines were adopted, the RWQCB would also need to make a policy decision on establishing new ceiling limits for foundation material.

The suggested values should not be used as a static list of screening guidelines. At every step of guideline development, decisions were made at critical junctions that, in practice, should be based on policy. Any change to these policy decisions would dramatically affect the results of the performance evaluations as well as the final recommended guideline numbers. The thresholds that can be altered in future re-analyses include the tolerance limits for assigning toxicity classifications, as well as the false negative and false positive rates considered acceptable for the purposes of wetland creation.

The suggested revisions to the sediment screening guidelines also required a revision to the management decision framework for evaluating material for wetland use. Figure ES-2 shows a proposed tiered testing framework that matches the level of required testing to the level of environmental protection associated with reuse of dredged material for either wetland projects (surface or foundation material) or landfill allocation. Some of the major differences in the framework shown above as compared with earlier versions include:

- Surface material guidelines act as true guidelines; if none are exceeded, the material is suitable for wetland cover with no required additional testing (as long as bioaccumulation triggers are not exceeded).

- The potential for bioaccumulation is recognized as an integral early decision tier for surface material considerations in this framework; however regional bioaccumulation triggers have not been established as yet and should be a high priority for the DMMO agencies.

- Permit applicants are given a choice (the "dredger's option") to subject the material to additional bioeffects testing instead of accepting the uncertainty associated with bioaccumulation trigger (BT) values or material with concentrations above surface SSGs.

The results from this project allow regulators and permit applicants to evaluate the suitability of dredged material for various disposal/reuse alternatives based on SSGs that were calculated on regional data and that have a known reliability performance. The proposed tiered testing framework, in addition to the proposed guidelines and the database and statistical methods developed to update these guidelines, can provide important tools for resource managers to reach LTMS goals for beneficial reuse.

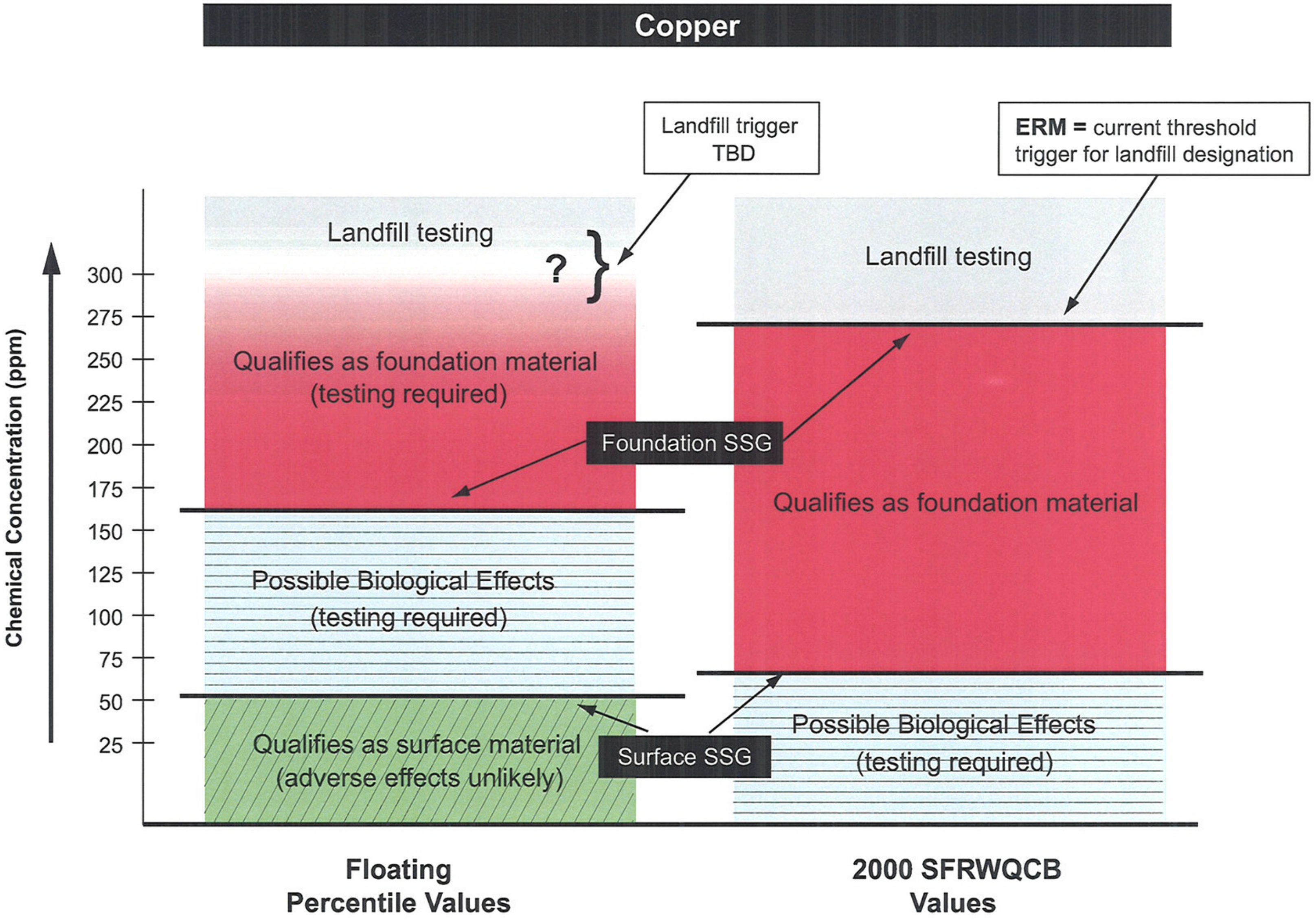


Figure ES-1. Differences in how SSG thresholds are defined in this report vs. the SFRWQCB 2000 report

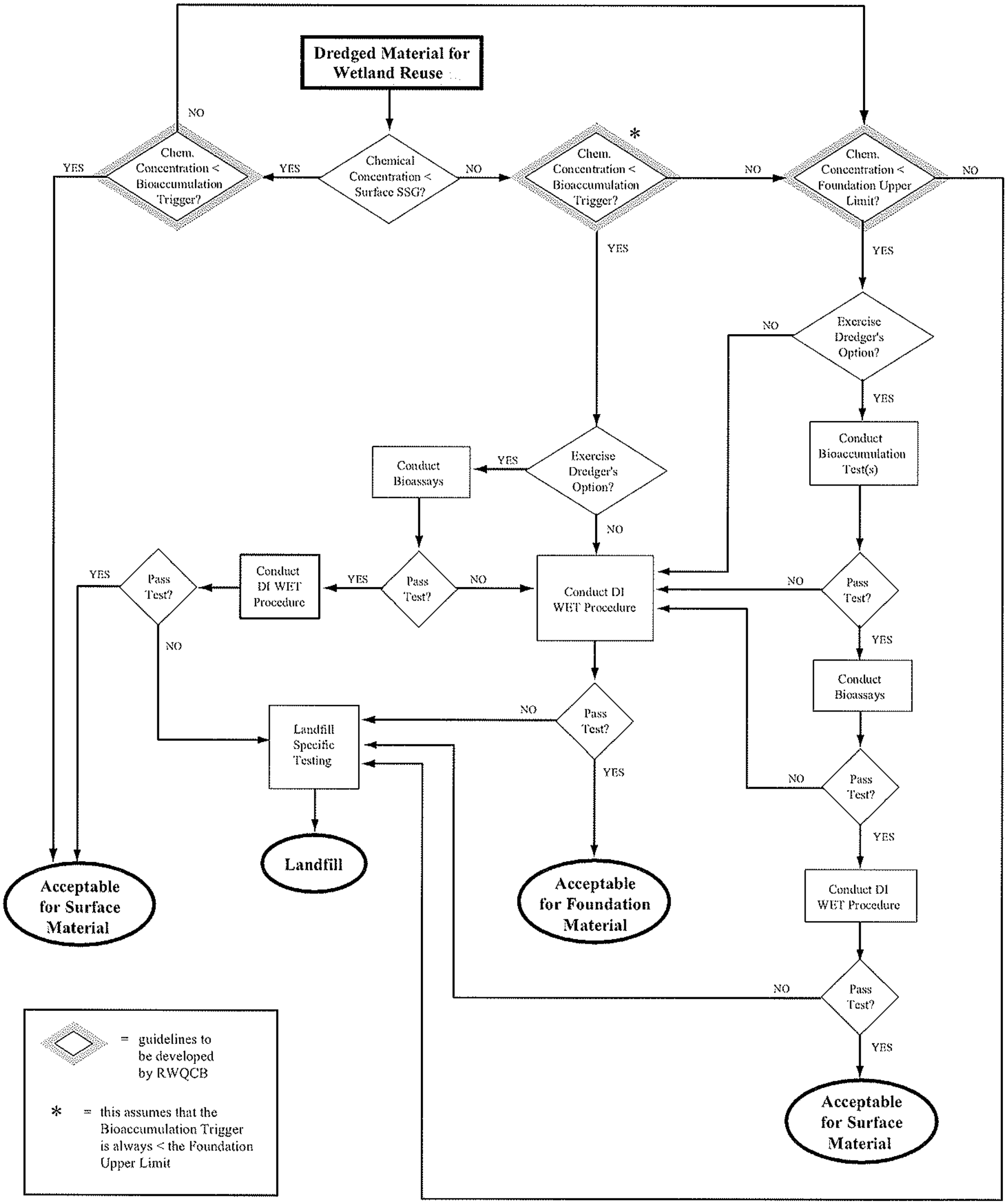

Dredged Material for Wetland Reuse
NO
YES
Chem. Concentration < Bioaccumulation Trigger?
YES
Chemical Concentration < Surface SSG?
NO
*
Chem. Concentration < Bioaccumulation Trigger?
NO
Chem. Concentration < Foundation Upper Limit?
NO
YES
YES
NO
Exercise Dredger's Option?
YES
Conduct Bioaccumulation Test(s)
Conduct Bioassays
YES
Exercise Dredger's Option?
NO
YES
Pass Test?
Conduct DI WET Procedure
YES
Pass Test?
NO
Conduct DI WET Procedure
NO
Pass Test?
NO
YES
Conduct Bioassays
NO
Landfill Specific Testing
NO
Pass Test?
YES
NO
Pass Test?
YES
Landfill
Acceptable for Surface Material
Acceptable for Foundation Material
Conduct DI WET Procedure
NO
Pass Test?
YES
Acceptable for Surface Material
= guidelines to be developed by RWQCB
* = this assumes that the Bioaccumulation Trigger is always < the Foundation Upper Limit

Figure ES-2. Proposed tiered testing framework for dredged material reuse or disposal

1.0 INTRODUCTION

This report summarizes the efforts over the past year to provide San Francisco Bay area regulators with an objective evaluation of sediment screening guidelines (SSGs) for dredged material used in wetland creation/restoration projects. This effort led to the suggested update of existing guidelines using site-specific sediment chemistry and bioassay testing results for the San Francisco Bay region. A series of interim technical memoranda and white papers that contain much of the detailed calculations and methods has been submitted to the project sponsors (California Coastal Conservancy, the regulatory agencies of the Dredged Material Management Office [DMMO], and the Port of Oakland) during the course of the project. These individual deliverables are provided in the appendices to this report. Also associated with this project was the creation of a structured Sediment Screening Guideline Database (SSGD) in Microsoft® Access 2000 to support the various evaluations and calculations done during the course of the program; a final version of this database, along with an updated design document and explanation of analysis routines, is being provided as a separate deliverable with this report. By having the database associated with the report, the project sponsors will be able to update the database in the future as more data become available, or to recalculate the guideline performance evaluations if they choose different reference tolerance intervals (Appendix B, Task 4.1 deliverable) or different specificity or sensitivity thresholds (Appendix C, Task 5 deliverable) for the newly proposed SSGs presented at the end of the report.

It is to the advantage of all regulators and stakeholders to have a performance evaluation of existing regional and national guidelines using current regional data to determine if these chemical thresholds are appropriate for the Bay Area. First we assembled existing regional paired data sets of both analytical chemical and biological effects testing results and then evaluated the reliability and accuracy of existing SSGs (both regional and national) at predicting the outcome of the paired bioeffects testing results. The results of this effort would provide an objective assessment of whether any set of existing guidelines performed optimally. If not, the assembled database could be used to derive new guidelines with more optimal performance for the regional results.

Because the ultimate decision on the suitability of dredged sediments for reuse rests on whether they cause adverse effects to biological receptors, an SSG is a useful resource management tool only if it can predict within a reasonable level of confidence the potential for a sediment to cause adverse biological effects. The main advantages of having reliable SSGs would be to: a) streamline the permitting process by giving both regulators and permit applicants a clear roadmap about which tests would need to be performed on sediments to be dredged, and b) eliminate the need to subject all sediments to the more costly suite of bioeffects testing (such as bioassay and bioaccumulation tests).

In the sections that follow, we present a brief background and rationale for revising the existing SSGs, provide an overview of the database structure (more detail can be found in the separate database deliverable), summarize the path traveled to complete the performance evaluations, explain how the new suggested guidelines were developed using the Floating Percentile calculation method, and finally discuss the results and their associated resource management implications for future wetland creation/restoration projects.

1.1 Background

Since the Long Term Management Strategy Environmental Impact Statement/ Environmental Impact Report (USACE, 1998) and the Record of Decision, published the following year, identified the "40-40-20" strategy to reduce disposal of dredged material in San Francisco Bay (40% allocated for the San Francisco Deep Ocean Disposal Site [SF DODS], 40% for beneficial reuse, and 20% for in-bay disposal), there has been increased interest in using dredged material as a resource for wetland restoration, levee repair, and landfill daily cover. Given the past history of wetland destruction associated with land reclamation in the Bay during this past century, wetland restoration has been a primary focus of state and regional resource management agencies. Of the Bay's original 2,200 km^2 of tidal saltmarsh, only about 125 km^2 remains (Nichols et al., 1986). Consequently, wetland restoration is an ideal alternative for dredged material beneficial reuse.

The Regional Water Quality Control Board (RWQCB) has published two sets of sediment screening guidelines for beneficial reuse of dredged material for wetland restoration in the past decade (SFB-RWQCB, 1992, 2000); however many of the guidelines defaulted either to ambient concentrations or to generic SSGs (Effects Range-Low, or ER-L, and Effects Range-Medium, or ER-M) of Long et al. (1995), which were derived from a national database. Recent publications have pointed out shortcomings with many of the national SSGs being used (Chapman, 2000; O'Connor and Paul, 2000) and how they were derived (Germano, 1999) In addition, local investigators have noted the lack of predictive reliability of these guidelines when applied to sediment chemistry results from dredged material characterization programs in the Bay Area.

In the two previous guideline documents issued by the RWQCB (SFB-RWQCB, 1992, 2000) on using dredged material for creation or restoration of tidal wetlands, a distinction was made between material earmarked for wetland surface use versus that designated for the underlying foundation sediments. Any dredged material used for the surface layer of wetland creation or restoration projects would be an integral part of the biotic zone and in contact with both flora and fauna; therefore, any sediment guidelines for surface material should be protective of sensitive biological receptors. The minimum thickness recommended for a surface layer is 3 ft, and project proponents are encouraged to maximize surface material thickness (SFB-RWQCB, 2000). Dredged material earmarked for foundation use (deeper than 3 ft), on the other hand, would be isolated from biological receptors, therefore the restrictions on allowable chemical concentrations are not as severe as those for surface material. The main concern for potential adverse effects from material used for wetland foundation would be contaminant mobility to either groundwater or biological receptors via leachate from foundation sediments once construction is completed; during construction, concerns with foundation material placement would include effluent run-off quality and prolonged exposure of foundation material to potential biological receptors.

Existing SSGs for surface wetland material in the RWQCB's most recent publication (SFB-RWQCB, 2000) were based primarily on ambient contaminant concentrations of sediments collected by the Regional Monitoring Program for Trace Substances and the Bay Protection and Toxic Substances Cleanup Program Reference Study. The ambient concentrations were calculated based on the 85th percentile with a tolerance confidence interval (confidence interval around the threshold percentile) of alpha = 0.05 (Gandesbery et al. 1998). The values were evaluated and adjusted relative to the grain size distribution; the final ambient values are recommended for fine-grained (100%) sediments. Ambient sediment concentrations were chosen by the RWQCB as the upper screening values for wetland surface material for two reasons:

1. Ambient values were generally lower than ER-L values, so it was assumed that they would be unlikely to cause adverse biological effects; when ambient values exceeded ER-L values (such as with nickel and chromium), past testing results of dredged material characterization programs have shown these elevated values of nickel and chromium concentrations have not been associated with adverse biological effects.

2. Because any restored tidal wetland would eventually take on the characteristics of the ambient sediments in nearby areas of the open bay, having screening values that would be lower than ambient values would be a waste of resources.

If sediments with concentrations of nickel and chromium higher than ER-L values have been shown to have no adverse biological effects in San Francisco Bay, then it stands to reason that concentrations of other contaminants higher than ER-L (or even ambient) values might also be benign, that is, have no adverse biological effects. Therefore, if SSGs for wetland surface material had higher values than ambient bay concentrations and were not associated with adverse biological effects, then surface sediments screened with SSGs higher than ambient values in any restored wetlands also would eventually "take on the characteristics of the ambient sediments in nearby areas" of San Francisco Bay.

1.2 Agency and Stakeholder Meetings

A series of meetings was held with regulatory agency representatives on September 11 and October 18, 2002, with an additional meeting on November 14, 2002, for both regulatory agency representatives and public stakeholders to present the database structure and our analysis approach, as well as to answer questions and get feedback on any concerns that participants may have had. Our first deliverable (Appendix A) presented a review and explanation of the various SSGs available and how they could be applied to evaluate sediments for wetland restoration projects in the Bay Area. There were a number of inherent limitations and assumptions that have been emphasized throughout the course of this project about the development of any SSG, and it is worth repeating some of them here.

First, any chemical-specific SSG, regardless of which approach was used for its development (Appendix A), will not be able to address the effects of unanticipated (or unanalyzed) chemicals that may be present in the sediment. Even though SSGs are usually developed for specific chemicals (for example, an ER-L for lead), SSGs will not always produce consistent results because sediments always contain mixtures of contaminants, and chemical-specific SSGs cannot address the potential synergistic or antagonistic interactions of various combinations of chemicals.

Some of the earliest regional SSG values developed were the Apparent Effects Thresholds (AETs) for the Puget Sound area (Tetra Tech, 1986), and even though attempts were made to use these Puget Sound AET values as SSGs in other settings (both domestic and international), investigators learned over the years that SSGs do not necessarily bear any relevance for geographic regions or environments other than those for which they were developed. An obvious corollary to this limitation is that any SSGs developed for this project based on San Francisco Bay data would not necessarily be relevant for wetland restoration projects in any region of the country (however, what is relevant and could be repeated elsewhere is the method and approach used to develop these regional guidelines).

Second, a concern that was raised at both the regulatory and public stakeholder meetings was that our entire evaluation as well as any suggested new guidelines developed from the data in our database would be based on marine amphipod acute endpoints; a justifiable criticism to this approach is that it would be more appropriate to base SSGs for wetland restoration projects on data from wetland receptors instead of subtidal marine organisms. One of the main reasons that all the regulatory guidance in the San Francisco Bay area on beneficial reuse of dredged material for wetland restoration up to this point has relied on marine amphipod tests as the ultimate pass/fail criterion for determining suitability of wetland surface material (SFB-RWQCB, 1992, 2000; USACE/USEPA, 1999a) was to take advantage of the data required for evaluation of in-Bay disposal. Even though marine amphipod bioassay results have been the qualifying factor for surface material acceptance in past regulatory frameworks, we are not merely arguing that maintaining the status quo is adequate justification for continuing the practice. There are several very practical reasons for using marine amphipod data to evaluate existing guidelines as well as to develop new site-specific guidelines:

The bulk of the data from which most of the national guidelines were derived (ER-Ls/ER-Ms, Threshold Effects Levels [TELs]/Probable Effects Levels [PELs], and Logistic Regression Models [LRMs]) was amphipod test results, so the one thing that these guidelines should be good at predicting is the outcome of an amphipod bioeffects test (for a short description of the amphipod test, see Section 2.1). Approximately 75% of the paired data in the regional database assembled for this project was amphipod test results, so it was the only data set with a sufficient number of samples to be able to perform the analyses required for this project.

There are no established bioeffects test procedures with wetland receptors, so no data are available (nationally or regionally) from which guidelines could be developed or evaluations performed. While other taxa from the order Amphipoda are present in wetland settings, the relevant question is not whether the subtidal genera used in most bioassay tests (e.g., *Ampelisca abdita, Eohaustorius estuarius,* and *Rhepoxynius abronius*) are found in wetlands, but whether they are valid surrogates for assessing potential adverse effects of the sediments under consideration.

Other than the potential for mercury methylation, the real issue is whether there is any basis for thinking that a particular sediment would be more or less toxic if it were placed in a subtidal versus a wetland setting. Given all of the above discussion, we feel that having the outcome from this project based primarily on amphipod bioassay results is actually the best available option for our results and conclusions.

The final set of major limitations and assumptions are based on some questions and concerns that were voiced in the public stakeholder workshop on November 14, 2002, and the responses provided to the workshop participants bear mention again for the readers of this report. While the database contains results from elutriate tests (see Appendix B), these results were not included in any of the performance evaluations or used for guideline development calculations, because they are not appropriate for assessing adverse sediment effects; elutriate tests are designed to mimic impacts to the water column at dredged material disposal sites during disposal operations, not to assess effects for receptors on or in a sediment substratum. The other important limitation to bear in mind when applying the suggested guidelines developed in this project is that bioaccumulation potential or effects have not been considered with the data in our database. The need to assess bioaccumulation potential is recognized and factored into the overall sediment management framework for wetland restoration (see Section 6.0, Conclusions and Recommendations), but the guidelines suggested in this report do not reflect either action thresholds or safe limits for chronic effects.

2.0 DATABASE STRUCTURE

2.1 Overview and Data Screening

Three databases were developed for this project. The main database (BayAreaSSG_MainDatabase.mdb) contains all of the sediment chemistry and toxicity data. Two other databases are linked to the main database, with imbedded queries developed for guideline analyses, including performance metrics (BayAreaSSG_PerformanceMetrics.mdb) and Receiver Operator Characteristics (ROC) analyses (ROC_Analyses_Final.mdb). Complete documentation of the main database is available in the user guide delivered with this report.

The database includes both dredging characterization data and regional monitoring data (Figure 1). The list of studies included in the database and the number of samples with at least one chemistry result and/or bioassay test, are shown in Table 1.

Table 1. Database study list

Study ID	Study Name	Agency	Type[1]	Publication Information			Number of Samples[2]
				Report Title	Report Authors	Year	
01	Port of Oakland 50t-ft Harbor Deepening	Port of Oakland	D	Comprehensive Final Sediment Analyses Report	EVS Environment Consultants, Inc	1998	86
02	SF Airport Sediment Characterization	SF Regional Water Quality Control Board	D	Phase II Sed Char Program Proposed Airfield Rconfiguration at San Francisco International Airport	Kinnetic Laboratories/ToxScan, Inc.	2001	72
03-04	SFEI 1993-94 RMP	SFEI	M	Regional Monitoring Program	SFEI	1993-94	87
05	BPTCP 1994/95 Reference Site Survey	Bay Protection and Toxics Control Program	M	Evaluation of Sediment Toxicity Tests and Reference Sites in San Francisco Bay - Draft	The Institute of Marine Sciences, University of California, Santa Cruz	1995	56
07	BPTCP 1995 Screening	Bay Protection and Toxics Control Program	M	Data report for Legs 38, 39, 40 and 41 submitted to SWRCB June 1996	State Water Resources Control Board	1996	97
08-11	SFEI 1995-98 RMP	SFEI		Regional Monitoring Program	SFEI	1995-98	211
DODS	SF-Deep Ocean Disposal Site	EPA Region Nine	D	SF-DODS Reference Area Database	Brian Ross, USEPA		6
J0	BPTCP 1996 Screening	Bay Protection and Toxics Control Program	M	Data Reports for Legs 38-42 [54]	State Water Resources Control Board	1996	10
Q0	BPTCP 1997 Confirmation	Bay Protection and Toxics Control Program	M	Data reports for Legs 47 through 56	State Water Resources Control Board	1998	30
RMP99, RMP00	SFEI 1999, 2000 RMP		M	Regional Monitoring Program	SFEI	1999-00	85
SFOBB	SFOBB East Span Project	California Department of Transportation (CalTrans)	D	SFOBB-East Span Seismic Safety Project	CalTrans and GeoCon	2000	19
SP	San Pablo Reference Area Data	Army Corps of Engineers	D	San Pablo Bay Reference Area Database	Jim Delorey, USACE	1989-97	5
U0	BPTCP 1997 Screening (Stege Marsh)	Bay Protection and Toxics Control Program	M	Data Reports for Legs 47 through 56	State Water Resources Control Board	1998	9
URS	Evaluation of US Steel Sediments	SF Regional Water Quality Control Board	D	Toxicity Evaluation of Sediments from the US Steel West Bay Cove Area	Pacific EcoRisk	2000	16

[1] D= dredging type study; M = monitoring type study

[2] Number of samples with at least one chemistry and/or bioassay result

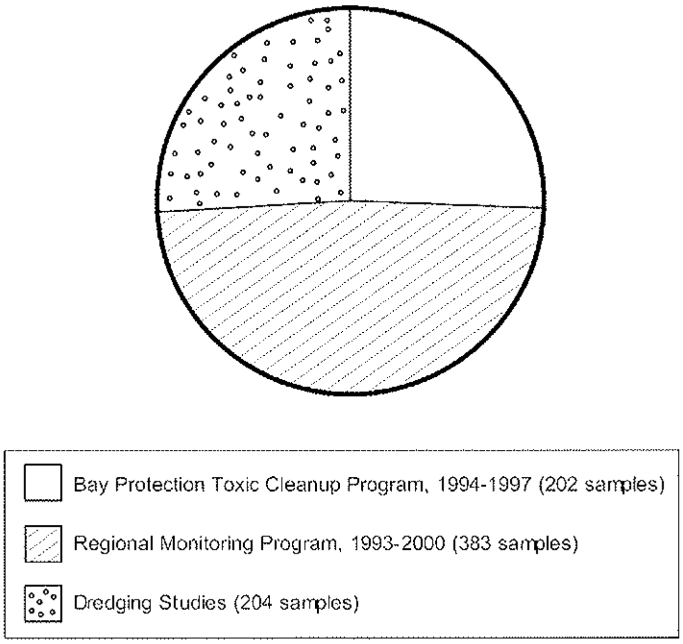


Figure 1. Samples in the database, by study type

The guidelines presented in this document are primarily based on the amphipod test, so a short description of this test is provided here. The amphipod test is a benthic bioassay designed to determine the potential impact of contaminated sediment on benthic organisms (ASTM 1998). Infaunal amphipods (Figure 2) are considered appropriate species for acute toxicity bioassays, because they are sensitive to benthic impact, readily available, and tolerant of a wide range of grain sizes and laboratory exposure conditions. Tests are conducted in aquaria, with strict controls on water renewal and water quality measurements. Bioassay tests should include a control sample and one or more reference samples; generally five replicates tests are performed for each sample. The standard test duration for acute amphipod toxicity bioassays is 10 days. Prior to the test, collected test organisms are observed to ensure they are healthy and have not been mishandled. A standard number of organisms are counted and placed in the aquaria for the 10 day period. After the exposure period, the sediment is siphoned through a fine mesh screen, and the animals are inspected. Amphipods that show any response to gentle probing are considered alive; those not recovered at the end of the test are considered as dead. If the control sample results in > 10% mean mortality (mean of the replicates), the test should be repeated. Unacceptably high control mortality indicates that the organisms are being affected by stresses other than contamination in the sediment. These stresses may be due to injury or disease, unfavorable physical or chemical conditions in the test containers, improper handling or acclimation, or possibly unsuitable sediment grain size.

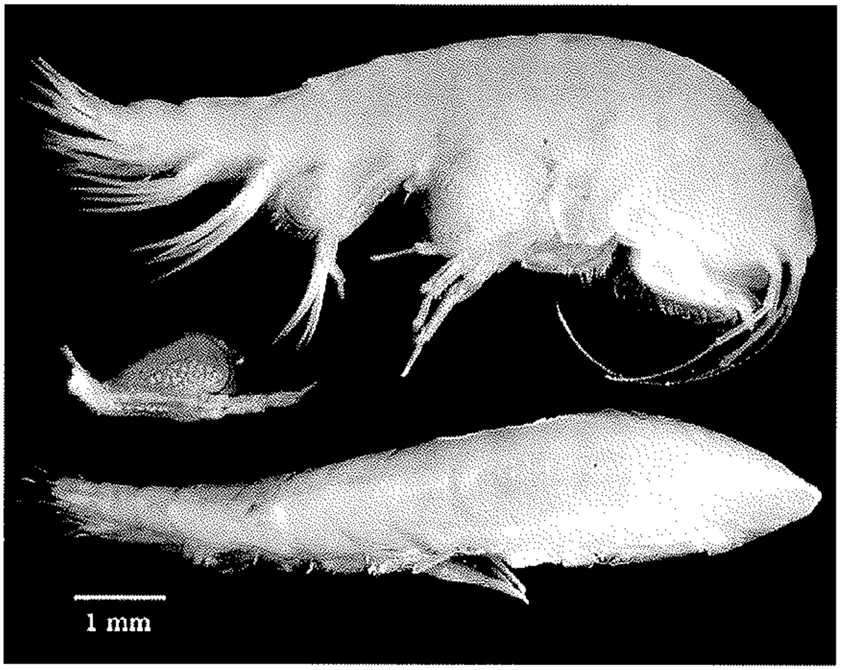


Figure 2. *Ampelisca abdita*, one of the amphipods commonly used for sediment bioassay tests.

For each individual sample, there are generally multiple sediment and elutriate bioassay tests (Table 2). During screening for the project, we identified the number of paired sediment chemistry/bioassay samples, a summary of bioassay species analyzed, and whether the data set contained total organic carbon (TOC) and grain size data. Each data set was evaluated to ensure that it was appropriate to be included for sediment guideline analysis. The screening criteria included:

- Well-documented quality assurance/quality control (QA/QC) information for both sediment chemistry and bioassay data for each study
- Accurate geographic coordinates
- Availability of grain size and TOC data
- Availability of toxicity replicate data
- Appropriate level of quality control for bioassay data, including meeting the minimum standard for negative controls for bioassay data (90%; ASTM, 1998)
- Reported toxicity significance and statistical methods used for mean toxicity data

Table 2. Number of unique samples for reference tolerance and test samples with synoptic sediment chemistry[1] and bioassay results by sediment sample matrix and bioassay species

	Reference Stations		Test Stations		Chronic Endpoint[2]	
Species-Endpoint	**Elutriate[3]**	**Sediment**	**Elutriate[3]**	**Sediment**	**Reference**	**Test**
Amphipod Mortality						
Ampelisca abdita		30		142		
Eohaustorius estuarius		69		245		
Rhepoxynius abronius		3				
Total unique samples		**64**		**337**		
Larval Mortality						
Mussels						
Mytilus edulis	1		57		18	143
Mytilus galloprovincialis			53		4	78
Sea Urchins						
Lytechinus pictus	2				2	
Strongylocentrotus purpuratus			13		2	25
Fish						
Citharichthys stigmaeus			12			
Menidia beryllina			50			
Total unique samples	**3**		**123**		**24**	**234**
Mysid Mortality						
Mysidopsis bahia		2	53	50		
Total unique samples		**2**	**53**	**50**		
Polychaete Mortality[4]						
Neanthes arenaceodentata		4				
Nebalia pugettensis		3				
Nephtys caecoides		4		62		
Total unique samples		**8**		**62**		

[1] At least one chemical in each sample is present in at least one guideline.

[2] The chronic endpoint of normality consists of all elutriate samples except 37 *S. purpuratus* sediment samples.

[3] The count of elutriate samples does not include individual dilution runs.

[4] The database also contains 4 samples of *N. arenaceodentata* growth.

2.2 Database Structure

The main sediment screening guideline database (SSGD) structure contains four levels of organization: Study, Station, Sampling, and Data (Figure 3). This organization reflects the very different sample design between dredging characterization and monitoring data.

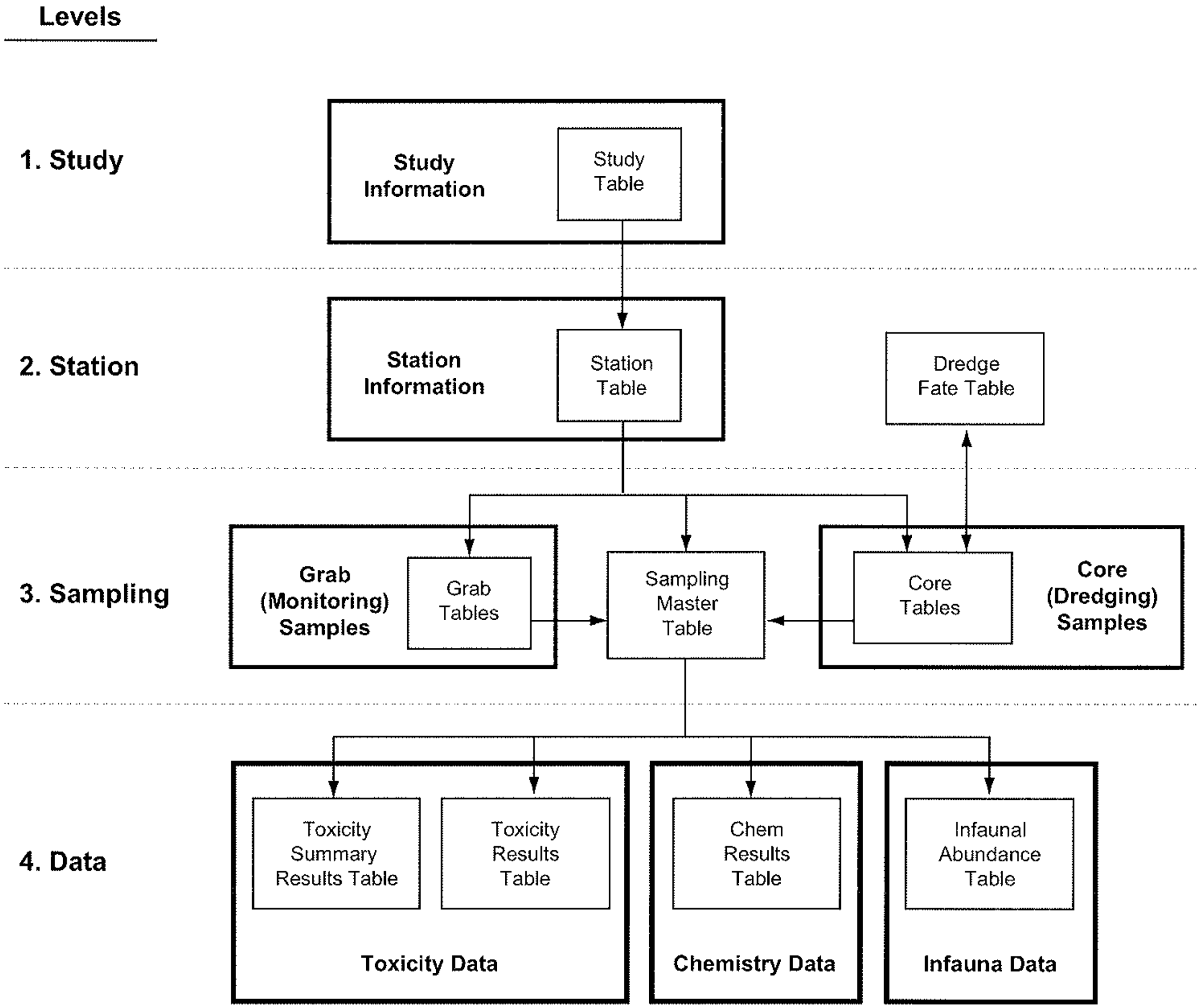


Figure 3. Organization of the Sediment Screening Guideline Database

The top-level hierarchy is the Study information. Every dredging report, as well as every monitoring data set, is one study. Each study has a unique identifier (*StudyID*) in the SSGD. The tables *tblStudy* and *tblStudyReference* contain information about each one of the studies in the SSGD.

The next level contains information about stations as well as dredging and monitoring data. Within that level, the database contains a series of tables that describe sampling information for the dredging and monitoring studies. Separate tables are used to document the sampling information for each type of study because of differences in study design and sample compositing among the methods. The final level of the database contains the data tables. These tables are organized by information type (e.g., chemistry, toxicity, or infauna) and contain the results of measurements from both dredging and monitoring studies.

All dredging and monitoring data have a geo-referenced location in latitude/longitude coordinates (NAD83). For dredging data, the Station may actually represent an area, such as a dredging polygon from which multiple cores were collected.

In the database, an additional field was added called *StationType* in the station table. This table was used to classify the reference stations that were used to calculate tolerance limits for standardization of toxicity status of test sediments. In addition to REF (reference stations), other types of stations are TEST (normal test station), CONT (control station), and OS (offshore or out-of-bay stations). Figure 4 shows the distribution of station types in the main database.

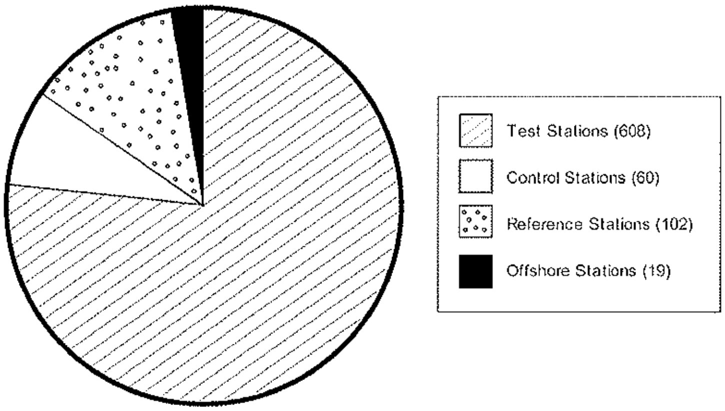


Figure 4. Distribution of station types in the main database

For toxicity data, the *Summary* table stores a series of summary values describing the results of that test. These fields include:

- Mean – mean value of laboratory replicates
- N – number of replicates

- StdDev – standard deviation of replicates
- PctControl – mean value expressed as a percent of the negative control assigned to that batch of samples
- SigEffect – reported statistical significance from original report and/or database
- NormSigEffect and NormSigEffect2 – standardized statistical significance developed for this project
- Stat Test – test used to calculate statistical significance
- LC50 – the concentration (%) of the sample that is lethal to 50% of the test organisms (applicable only to the endpoint of survival and usually only reported for dredged material elutriate tests)
- EC50 – the concentration (%) of the sample that produces an adverse effect on 50% of the test organisms (applicable to sublethal endpoints and usually only reported for dredged material elutriate tests)

The codes used for statistical significance for the *SigEffect* were derived from the original report, and differentiate between comparisons to reference and control samples. There are also two additional fields for significance calculated from the reference tolerance threshold application. The codes used for statistical significance for the *NormSigEffect* field were developed only for solid phase amphipod acute toxicity data, based on a Reference Envelope approach (see Section 3.1). A sample that had a mean survival that was less than the Reference Envelope tolerance limit would be considered toxic. Selection of a tolerance limit based on the 10th percentile is equivalent to calling a sample toxic if it is expected that its performance is worse than the performance of 90% of the reference samples. An example of how reference envelope tolerance limits are computed and applied is presented in Hunt et al. (1998) and CSWRCB (1998). This definition was later altered to represent the 20th percentile for *E. estuarius* data, as described in Appendix B. This second definition is stored as the field *NormSigEffect2*.

Finally, the additional table *TblToxRefSigEffect* contains the reported values for significance compared to multiple reference areas for one study (SFOBB East Span Project; Table 1). One other study had multiple reference comparisons (San Francisco Airport Sediment Characterization), but the results were the same for all reference areas.

The main database contains a table called *SQG_FinalTable* that has all guideline values evaluated, including the final site-specific guideline numbers (*FPSurface* and *FPFoundation*). Note in this table that each guideline is reported with a Unit field (dry weight [DW], organic-carbon normalized, or molar). The values are as published, except for nickel and chromium which were elevated to background concentrations (SFB-RWQCB, 2000).

2.3 Data Analysis Content and Programs

Data analysis routines are stored in two separate databases. For efficiency, the analysis databases include links to the data tables in the main database. These links will need to be refreshed to your local drive should you choose to re-run these analyses; see the database User Guide for further information.

For toxicity data, the final guidelines were developed using only amphipod survival for *E. estuarius* and *A. abdita,* excluding reference, control, and offshore stations (SF-DODS, Tomales Bay, and Bolinas). Both individual and pooled amphipod species were used. Additional analyses were conducted using other sediment endpoints (*Nephtys caecoides* and *Mysidopsis bahia*); queries for the pooled sediment test endpoints are available in the database as well.

For chemistry data, a *UseResult* field was first created, representing the results for most records; for values reported as below detection limits, ½ the detection limit was entered instead (except for the Floating Percentile analyses, which excluded data below detection). In most cases, total polycyclic aromatic hydrocarbons (PAHs), DDTs, chlordane, and hexachlorocyclohexanes (BHCs) were recalculated using a standardized formula (see below). Missing data (reported as –99) were filtered out of the table, and laboratory replicates were averaged. The chemistry data were then linked with the toxicity data so that only amphipod acute toxicity samples were extracted. Then, for samples with available TOC data, a TOC-normalized value was created on a parts-per-million basis for non-ionizable organic compounds.

Although most of the performance analyses were calculated on individual chemical guidelines, we also investigated the performance of the mean sediment quality guideline quotient, referred to here as the SQG-Q1 (Fairey et al. 2001). The SQG-Q1 is a mean quotient calculated from nine chemicals that, in combination, best predicted amphipod toxicity (Cd, Cu, Ag, Pb, Zn, chlordane, dieldrin, total PAHs, and total PCBs). The quotient is calculated by dividing each measured concentration by a selected guideline, summing each normalized value, and then calculating a mean quotient for each sample. Fairey et al. (2001) used a combination of ER-Ms, PELs and consensus values for their final SQG-Q1 guidelines (definitions of the guidelines in Section 3.2 and Appendix A).

In general, only samples that had at least one chemical in one guideline were included in performance calculations. For the SQG-Q1 quotient analysis, only those samples with all nine SQG-Q1 chemicals and TOC measured were included in the quotient calculations. For the Floating Percentile analysis, only samples that had at least one metal and at least one PAH value were included. Because the Floating Percentile dataset excluded data below detection and included only samples with at least one metal and one PAH, it is described as a 'restricted' dataset in comparison to the 'complete' dataset used for the performance calculations in this discussion.

For summed parameters, the data were handled in the following manner:

- *PAHs, dry weight* – There are three PAH sums in the database (low molecular weight PAHs [LPAH], high molecular weight PAHs [HPAH], and total PAHs). Most are calculated sums from the individual PAHs. For one study (StudyID = "URS"), only total PAHs are in the database as reported, because there were no individual PAHs reported, or low or high molecular weight sums. PAHs were calculated excluding data below detection limits. The list of chemicals used to calculate PAH sums is available in the database documentation and Appendix C, Section 2.1.

- *PAHs, molar* – There is also a chemical name called total PAHs (molar), that provides a calculated sum of all of the PAHs normalized to their molecular weight (see Section 4).

- *PCBs* – Because total PCBs was reported variously as congeners and Aroclors (with a different number of compounds analyzed for each study), total PCBs was not standardized. The Port of Oakland and San Francisco Airport studies reported only total Aroclors; thus Aroclors were used for calculation for total PCBs. SFEI measured only congeners, and reported total PCBs as total congeners. The BPTCP and San Francisco Bay Bridge studies measured both, but total PCBs were calculated using congeners. The value of total PCBs used in all statistical analyses was the reported total.

- *DDTs, dry weight* – Currently in the chemical results table, the chemical name DDTs represents the total of any and all isomers, excluding all values below detection limits. The original reported total DDT value from each report is also available in the table *ReportedPAHsDDTs*.

- *Total Chlordane* – Chlordane was calculated as the sum of the chemicals shown in the database documentation. If all chlordane chemicals were reported as below detection limits, the highest detection limit of the chlordane chemicals was used for total chlordane.

- *Total BHCs* – Total BHCs were calculated using only detected BHC compounds from the sum of alpha-, beta-, delta-, and gamma-BHC (lindane).

The database *BayAreaSSG_PerformanceMetrics.mdb* contains the data, queries, and tables required to calculate performance metrics of individual guidelines. Two types of analyses were conducted testing two types of predictions:

- *Prediction Definition 1 (Single Exceedance)* – For this method, a sample was predicted to be toxic if one or more chemical guidelines were exceeded; most of the guidelines in the *SSG_FinalTable* used this method, except for the SQG-Q1 quotient value.

- *Prediction Definitions 2 and 3 (Mean Quotient Prediction)* – For this method, a sample was predicted to have a low probability of acute toxicity if the mean SQG quotient (rounded to 2 decimals) was less than or equal to a threshold. Prediction 2 used a quotient threshold of 0.10; Prediction 3 used a quotient threshold of 0.5.

The database *ROC_Analysis_Final.mdb* contains the data, queries, and tables required to conduct ROC analyses. There are two programs created for ROC analyses. These can be activated by opening the two forms "ROC Calculations" and "Run AUC" (AUC = area under the curve). They must be run in the following order:

- *ROC Calculations* – This program runs a series of queries that compares each reported value (as a potential guideline) for each chemical to the actual value. If the measured value is greater than the potential guideline value, then that sample is stored as a 'hit' for that potential guideline value. Performance metrics similar to the queries above are then run for the samples predicted to be toxic (one minus specificity, or the false positive rate, and sensitivity, or the true positive rate). The program is currently setup to use the table ChemData (excluding data below detection), and the pooled amphipod end point (20th percentile definition for *E. estuaries*). Chemicals without at least one hit and one no-hit are removed from the calculations and stored in the 'BadData' table. Final results are stored in the Results table. Note the program will take a few minutes to run, with a 'Done' message at the end. The results then can be imported into an Excel® file to plot the ROC curves.

- *Run AUC* – Once the values are generated that will form the ROC curves, a program called Run AUC was created to calculate the area under the curve using the trapezoid rule (adding up each trapezoid under a curve). It is not necessary to generate the curves (created in Excel) to calculate the AUCs. You have the choice of calculating the AUC for one selected chemical, or for all of them at once. This program should run quickly; thus no message is provided when complete. The output (stored in the table AUC_Final) will include both the AUC, and the number of samples used to calculate the output. The output of the program results in a list of chemical names and AUC values.

3.0 PERFORMANCE EVALUATIONS OF EXISTING SSGS

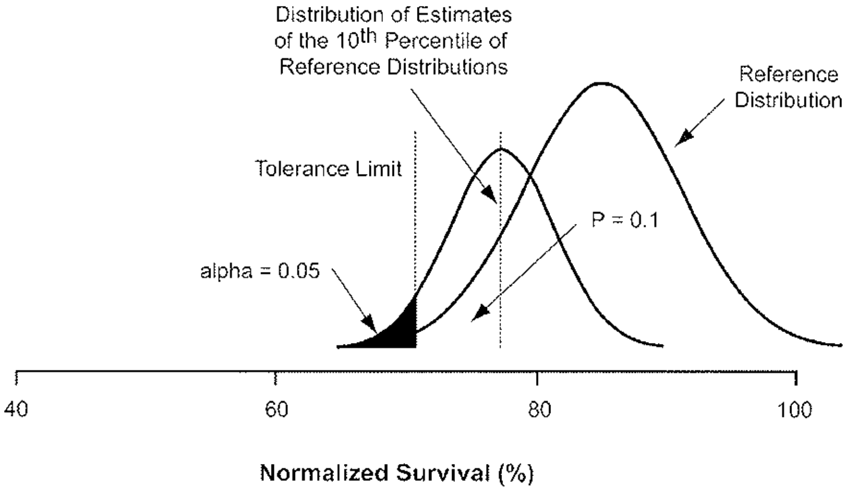
The synoptic sediment chemistry and bioassay results were used to evaluate the performance of several existing sets of SSGs. This evaluation involved predicting biological effects by comparing sediment chemical concentrations to the existing SSGs, and then comparing these predictions with the observed toxicity status for the amphipod mortality bioassay results.

3.1 Reference Envelope Tolerance Limit Calculations

Prior to the performance evaluation of existing SSGs, the acute toxicity status (i.e., toxic or non-toxic) of bioassay results for samples in the database were standardized using the Reference Envelope approach (summarized below and described in more detail in Appendix B). The dredging and monitoring studies in the database were originally assigned toxicity designations based on pair-wise statistical comparisons to either native control or batch-specific reference sediment results. Consequently, the mean survival across non-toxic or toxic samples varied considerably, depending on the response of the control or reference sample to which each test sample was compared.

The standard used to assign toxicity status to each test sample will have a substantial impact on the results of a performance evaluation. In some situations, comparisons to native control sediments have been shown to provide better reliability results than comparisons to reference samples (SAIC/Avocet, 2002). However, this is likely due to control sediment responses being more standardized than reference sediment responses, because the limits for acceptable native controls are tighter than the limits (if any) used to determine acceptable reference samples. Native control sediments are a laboratory QA measure used to verify the health of the animals through the experimental process. As such, they provide important information about the quality of the animals used in a particular bioassay test batch. The primary objective for wetland creation applications is to match ambient biological conditions in nearby open bay locations (Section 1.1). The question of how a test sample compares to local ambient conditions is addressed by basing toxicity status on reference responses.

The Reference Envelope approach uses the full range of biological responses observed from exposure to high-quality reference samples. A summary statistic derived from this distribution of reference responses serves as a threshold for determining which test samples are toxic. The statistic used in previous Reference Envelope applications in the Bay Area was a tolerance limit (Hunt et al., 1998). A tolerance limit is a confidence bound on a percentile of the underlying population. Our reference distribution is just a single sample from the population of all possible reference distributions. A percentile computed from our particular reference distribution is just one of many such percentile values possible, and so this percentile is itself a random variable with a distribution (Figure 5).



Note: In this figure, the tolerance limit is the lower 95% confidence bound ($\alpha = 0.05$) on the 10th percentile ($P = 0.1$) of the distribution of normalized survival values from reference area samples.

Figure 5. Graphical representation of a tolerance limit

The data preparations and tolerance limit calculations generally followed the same approach used by Hunt et al. (1998). Reference distributions were assembled separately for *A. abdita* and *E. estuarius* from the reference locations used in Hunt et al. (1998) and shown in Table 3. Sampling dates ranged from spring 1993 to summer 2000.

Table 3. Sampling locations and sample sizes for the candidate reference distributions used in tolerance limit calculations

Ampelisca abdita	*Eohaustorius estuarius*
North South Bay (n=4)	North South Bay (n=7)
South South Bay (n=3)	South South Bay (n=5)
Paradise Cove (n=12)	Paradise Cove (n=12)
San Pablo Bay Island #1 (n=9)	San Pablo Bay Island #1 (n=11)
Tubbs Island (n=11)	Tubbs Island (n=11)
	Pinole Point (n=4)
	San Bruno Shoal (n=13)
	Horseshoe Bay (n=13)
Total Candidate Samples = 39	Total Candidate Samples = 76

The candidate reference samples were screened using a strict bioassay QA standard. Reference samples from batches with native control sediments that failed the American Society for Testing and Materials guideline of a minimum 90% survival (ASTM, 1998) were excluded from the final reference distribution. Three test batches had native controls with less than 90% survival, resulting in the rejection of 3 of the 76 candidate *E. estuarius* reference samples and 18 of the 39 candidate *A. abdita* reference samples from the final distribution. The native control acceptance criterion was not implemented in the previous tolerance limit calculation, so 20 of these 21 rejected samples had been included in the original tolerance limit calculations done by Hunt et al. (1998).

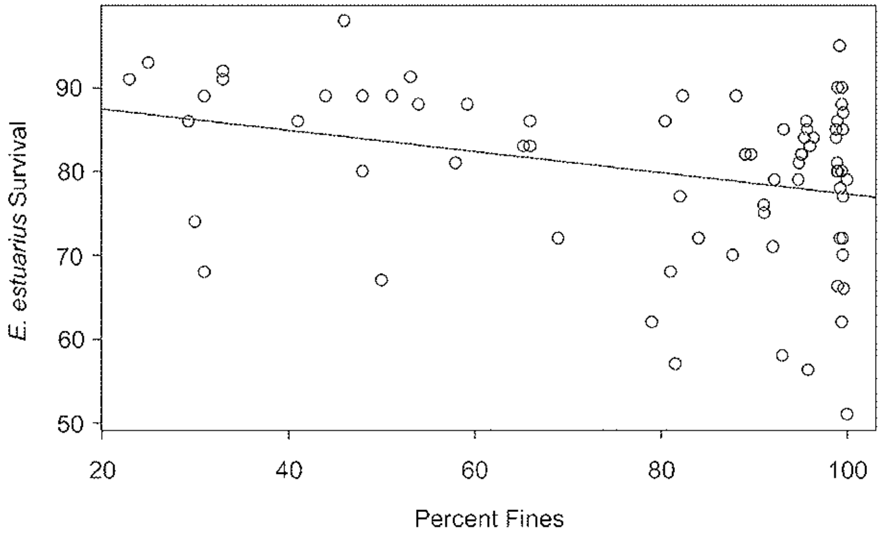
To ensure there were no contaminant concentration outliers in the sediments used to calculate this reference envelope, chemical concentrations in the candidate reference samples were evaluated using a mean SQG-quotient approach based on ER-Ms and PELs. The reference samples included in this study frequently had detected concentrations of metals and anthropogenic chemicals, but these concentrations tended to be well below the guideline values; mean quotient results were consistent with the results reported for other reference studies. Overall mean ER-M quotients for 17 to 22 individual substances (nickel and chromium were excluded from this mean because of high background concentrations in Bay area reference sediments) ranged from 0.04 to 0.16. Mean PEL quotients for 19 to 25 individual substances (nickel and chromium excluded) ranged from 0.09 to 0.33. Summary results for this and other studies reporting mean SQG quotients for California coastal reference areas are shown in Table 4. Details on the data collection, treatment, and biological and chemical screening results can be found in Appendix B, Sections 4.1 – 4.3.

Table 4. Summary of mean quotient results of California Coastal reference samples

Study	Range of Mean ER-M Quotients	Range of Mean PEL Quotients
This study	0.04 – 0.16	0.09 – 0.33
Hunt et al. (1998)	0.09 – 0.27	< 0.37
Fairey et al. (1996)	0.07 – 0.25	0.12 – 0.40

Physical characteristics (TOC and grain size) of the reference sediments were within the range of those found in the test sample database. The relationship between *Eohaustorius* sp. survival and grain size was investigated. The potential detrimental effects of sediments with high clay fractions on *Eohaustorius* sp. survival have been noted by several researchers (e.g., Dewitt et al., 1989; Hunt et al., 1998), and the potential for these effects have been incorporated into some regulatory bioassay testing requirements (USACE/USEPA/WDNR/WDOE, 2000). Unfortunately, the individual studies in this project database did not use the same methods for determining the sediment clay fraction, confounding any observable relationship between percent clay and *E. estuarius* survival (see Appendix B, Section 4.4 for details). The best alternative measurement for grain size effects in this data set is total fines (silt + clay fractions).

The relationship between *E. estuarius* survival and percent fines was found to be statistically significant (correlation coefficient = -0.31, p = 0.007; Figure 6). However, despite the statistical significance, this relationship is weak, with substantial scatter around the best fit regression line (Figure 6; Table 5).



Note: Correlation coefficient for this relationship is -0.31 (p = 0.007).

Figure 6. Scatterplot and best fit regression line (by ordinary least squares) between percent fines and *Eohaustorius estuarius* survival

Table 5. Summary of *E. estuarius* survival in two sediment types

Samples with:	Mean Survival (%)	Survival Range (%)	Sample Count
Fines > 80%	78	51 – 95	49
Fines < 40%	86	68 – 93	8

The presence of confounding factors in this data set (including different clay measurement methods as well as potential laboratory effects such as acclimation procedures and test organism sources), the weak correlation between *E. estuarius* survival and percent fines, and the highly variable and overlapping *E. estuarius* response for all sediment types make it untenable to correct for the effect of grain size on the *E. estuarius* responses in this particular data set. Consequently, the tolerance limits computed here are based on the reference distribution in its entirety. This tolerance limit will be generally applicable to all sediment types, but may result in a slightly lower threshold than might be expected for very sandy sediment types. More details on this issue can be found in Appendix B, Section 4.4.

The endpoint used in the tolerance limit calculations was normalized survival (i.e., the survival in the reference sample expressed as a percent of the native control survival). All figures and tables show the normalized survival values, unless otherwise noted.

The tolerance limit calculations used a parametric computational method (Bagui et al., 1996) with a double bootstrap calibration procedure (Efron and Tibshirani, 1993) to correct for bias in the sample estimates (required by this particular computational method; see Smith, 2002). Summary details for this application are found in Appendix B, Section 3.6; for a more complete discussion of the methods, see Smith (2002), Smith and Riege (1999), and Hunt et al. (1998). All computations and simulations in this application used the TIA computer program[1] developed by Robert Smith.

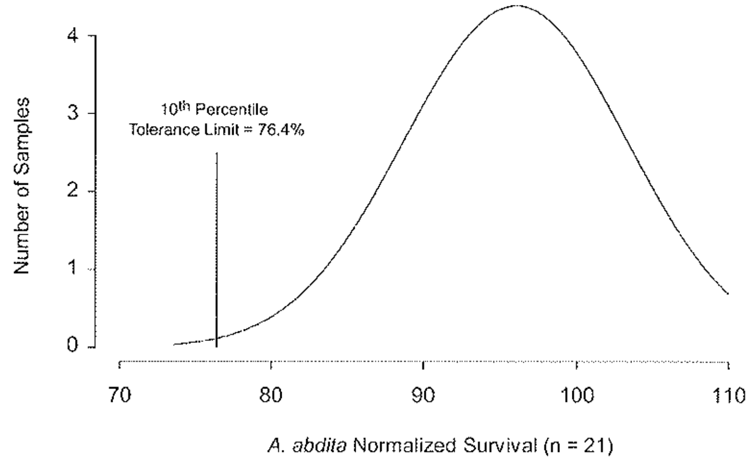
The parametric method used to compute the tolerance bounds requires that the data be normally distributed. Also, because the tolerance bounds of interest are in the tail of the distribution, the presence of outliers can be highly influential and need be removed prior to tolerance limit calculations. The *A. abdita* data were found to have no outliers and were approximately normally distributed. The *E. estuarius* data were also found to have no outliers but needed transformation to comply with the computational requirement for normally distributed data. Details on the outlier and goodness-of-fit tests for normality can be found in Appendix B, Section 4.5.

Tolerance limits based on approximate 95% confidence bounds on percentile values ranging from 1 to 25 were calculated for each species (Appendix Table B-6). The best success for SSG predictions is achieved in a data set with a clear delineation between chemically contaminated toxic sites and chemically uncontaminated non-toxic sites. For reference distributions with a high mean and low variance, a low percentile tolerance limit

[1] Development of this computer program was partially funded by California State Water Resources Control Board, EcoAnalysis Inc., EPA Region 9, and the San Francisco Bay Regional Water Quality Control Board.

(e.g., 10th percentile) may effectively delineate contaminated from uncontaminated sites. However, very low variance together with a high mean can produce a 10th percentile tolerance limit that is too high to be biologically meaningful. On the other hand, for reference distributions with a high variance and/or low mean, a higher percentile tolerance limit (e.g., 20th percentile) may be necessary to avoid an overly liberal acute toxicity threshold. Excessively high variance together with low mean values in the reference distribution can generate negative tolerance limits that are obviously impractical. The appropriate percentile to use for a tolerance limit depends on the application, as well as the confidence that the reference distribution adequately characterizes optimal (ambient) conditions. Following the general precedent set by Hunt et al. (1998), the tolerance limit based on the 10th percentile was initially considered appropriate.

For the *A. abdita* reference distribution, normalized survival averaged 96% and ranged from 82% to 108% (Figure 7; Appendix Table B-1). The sample size of 21 is at the lower end of the number of samples needed to adequately describe the reference distribution and compute bootstrapped estimates. The distribution itself is almost multi-modal (Figure 7). There is a significant location effect on the normalized survival response (a two-factor ANOVA had a p-value of 0.01 for the location effect and a p-value of 0.49 for the season effect), with Tubbs Island samples having the lowest survival responses (median normalized survival value is 85%). However, there is still quite a bit of overlap among normalized survival responses for the different reference locations and the sample sizes are quite small within each location (n=3 to 6). More observations should smooth out the distribution and provide more confidence in the estimates derived from it. We used an acute toxicity threshold based on the 10th percentile tolerance limit for *A. abdita* that corresponds to a normalized survival value of 76.4%; this resulted in 24 samples designated as toxic and 123 samples designated as non-toxic.



Note: The 10th percentile tolerance limit (α = 0.05) is shown.

Figure 7. Histogram of reference responses for *A. abdita* normalized survival with the probability density function overlaid for a normal distribution with the mean and variance observed for these data

This tolerance limit should be updated after the *A. abdita* reference distribution includes at least 30 or more high-quality reference responses. The previous value based on 34 samples (which included 18 samples that were excluded from this calculation due to poor native control performance) resulted in a value of 70.9% (Hunt et al., 1998).

For the *E. estuarius* reference distribution, normalized survival averaged 83% and ranged from 52% to 102% (Figure 8). The sample size of 73 is very good for characterizing the reference distribution, and the distribution itself is smooth with very few value gaps. The lowest survival value (52%) was part of the reference distribution in the previous tolerance limit calculations (Hunt et al., 1998) but was excluded from those final calculations because it was identified as an outlier. This same data point was not an outlier given the current data distribution, so this is an example of how more data (13 more samples than previously) filled in the gaps of the distribution and validated the result. This species appears to have a large amount of variability in the bioassay response. The high variance in this distribution may be attributed to a number of confounding factors (e.g., grain-size effects, acclimation procedures, etc.). However, the effect of these factors could not be adequately explained by the available information. A sensitivity analysis indicated that the 20th percentile tolerance limit provided a better separation between chemically contaminated and uncontaminated sites, thereby improving the performance of SSGs at predicting acute toxicity.

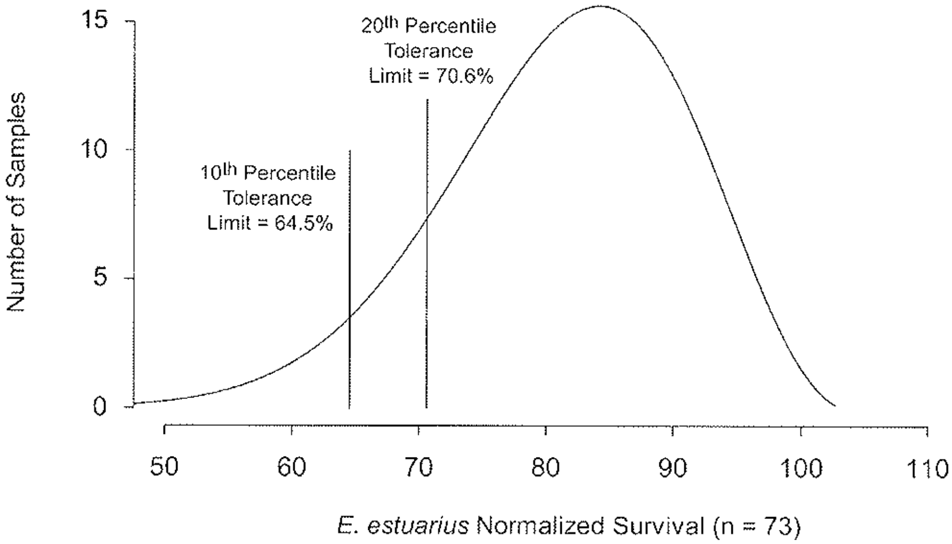


Figure 8. Histogram of reference responses for *E. estuarius* normalized survival with the probability density function overlaid for a normal distribution (on the arcsine-square root transformed scale) with the mean and variance observed for these data

As a result, we used an acute toxicity threshold based on the 20th percentile tolerance limit for *E. estuarius,* which corresponds to a normalized survival value of 70.6%; this resulted in 124 samples designated as toxic and 204 samples designated as non-toxic. The previous 20th percentile tolerance limit based on 60 samples was 73.4% and 69.5% for the 10th percentile (Hunt et al., 1998).

The choices of which percentile value to select from the reference distribution and the expected confidence level in that value are strictly policy decisions. Statistical methodology simply provides a way of estimating a summary statistic that meets the desired policy objectives. Applications of tolerance limits in the Bay Area have historically used tail-end percentiles: for example, the 10th percentile for reference bioassay results was used for establishing toxicity thresholds (Hunt et al. 1998), and the 85th percentile for ambient sediment concentrations were largely the basis for the 2000 beneficial reuse surface screening guidelines (Gandesbery et al. 1998). In addition, the chosen alpha level is consistent with the tendency in regulatory settings to fix the Type I error at 5%: examples of this include the 95% upper confidence bound on the mean in risk assessment guidance (USEPA, 1992a) as well as the 95% confidence levels in tolerance limits for ground water monitoring (USEPA, 1992b), reference toxicity thresholds (Hunt et al. 1998), and ambient sediment concentrations (Gandesbery et al. 1998).

The toxicity thresholds chosen for use in the remainder of this study provide 95% confidence that at least 90% (P=0.1) of the reference survival values will exceed the toxicity threshold for *A. abdita*; and 95% confidence that at least 80% (P=0.2) of the reference survival values will exceed the toxicity threshold for *E. estuarius*. For *Ampelisca* the Type I error is the probability that fewer than 90% of the *Ampelisca* reference survival values will exceed the toxicity threshold (i.e., that the 10th percentile from a reference distribution of *Ampelisca* results will be below the tolerance limit). Only site samples that have a survival value clearly different from the reference distribution will be labeled "toxic." A prerequisite for this or any approach to defining toxicity that uses statistical comparisons to reference is that we must be confident that the reference distribution includes only high quality bioassay results from uncontaminated sites reflecting ambient conditions. Uncertainty about the reference distribution can be accommodated by adjusting the targeted percentile in the tolerance limit. The higher percentile value for *E. estuarius* was chosen because of the high variance in this distribution. Note that alternative toxicity thresholds from the Reference Envelope approach are provided in Appendix B.

3.2 Reliability Results for Existing SSGs

Once the tolerance limits were established for the reference samples, we could then predict biological effects from sediment chemistry in the regional database using a number of existing SSG sets. The SSG sets included in this effort were selected because of their common use in national or other regional programs. While these SSG sets are based on several different models representing different probabilities for adverse biological effects, all of them are based in some way on an empirical correlative relationship between acute toxicity and bulk sediment chemistry. The SSG sets representing a low probability of acute toxicity were evaluated as potential screening guidelines for wetland surface material. Guideline sets representing higher probabilities of acute toxicity were evaluated as a screen for sediments to be used as wetland foundation material (which have no direct exposure route).

The sets of guideline values included in this evaluation are identified in Table 6. Additional SSG sets evaluated during preliminary runs and the guideline values for all compounds for all SSG sets used in this evaluation are shown in Appendix Tables C-2 and C-3. Each SSG model is described more fully in Appendix A.

Table 6. Existing sets of sediment screening guideline values evaluated against the San Francisco Bay database

SSG Values	Source	Comments
1. Wetland screening criteria in San Francisco Bay (surface)	SFB-RWQCB (2000)	Mainly ambient concentrations, with a couple ER-Ls. Dry weight. Not species specific.
2. Wetland screening criteria in San Francisco Bay (foundation)	SFB-RWQCB (2000)	Mainly ER-Ms with a couple of PELs. Dry weight. Not species specific.
3. ER-Ls (surface)	Long and MacDonald (1992)	The concentration below which biological effects are rarely observed. Dry weight. Not species specific.
4. TELs (surface)	MacDonald et al. (1995)	The upper limit of the range of concentrations that are not likely to be associated with adverse biological effects. Dry weight. Not species specific.
5. PELs (foundation)	MacDonald et al. (1995)	The lower limit of the range of concentrations that are usually associated with adverse biological effects. Dry weight. Not species specific.
6. Logistic Regression Models (LRM) T20 values (surface)	Field et al. (2002)	Individual chemical guidelines derived from a concentration-response curve. Each T20 value is the concentration associated with a 20% probability of toxicity. Dry weight. Based on mortality for amphipod test species *A. abdita* (60% of data) and *R. abronius* (40%).
7. LRM T40 values (foundation)	Field et al. (2002)	Same models used in SSG Set #6, just a different point along the curves. Each T40 value is the concentration associated with a 40% probability of toxicity.
8. LRM multi-chemical P_{max} 40% model (foundation)	Field et al. (2002)	Same models used in SSG Sets #6 and #7. Multi-chemical model associated with a 40% probability of toxicity adjusted for overestimation of toxicity probability (*see* text).
9. SQG-Q1 (used for both surface and foundation using different thresholds for the mean quotient)	Fairey et al. (2001)	The average of a set of 9 chemical guideline values that provided optimal reliability in several independent data sets. Dry weight, with one organic carbon normalized value. Not species specific.

These existing regional and national guideline sets were evaluated for their ability to accurately predict pooled amphipod acute toxicity or non-toxicity in the San Francisco database. The toxicity designation of the pooled amphipod endpoint was determined by whatever *A. abdita* or *E. estuarius* result was available for each sample. For samples that had both *A. abdita* and *E. estuarius* test results, the pooled endpoint was classified as toxic if either the *A. abdita* or *E. estuarius* result was classified toxic. Performance for individual amphipod species and for a pooled endpoint including *Mysid* and *Nephtys* is shown in

Appendix C. Results for these other endpoints were generally consistent with the results for the pooled amphipod endpoint reported here. The performance evaluation for each guideline set involved:

1. Predicting acute toxicity/non-toxicity for each sample based on sediment chemistry

2. Tallying samples in each of the four categories of predictions:

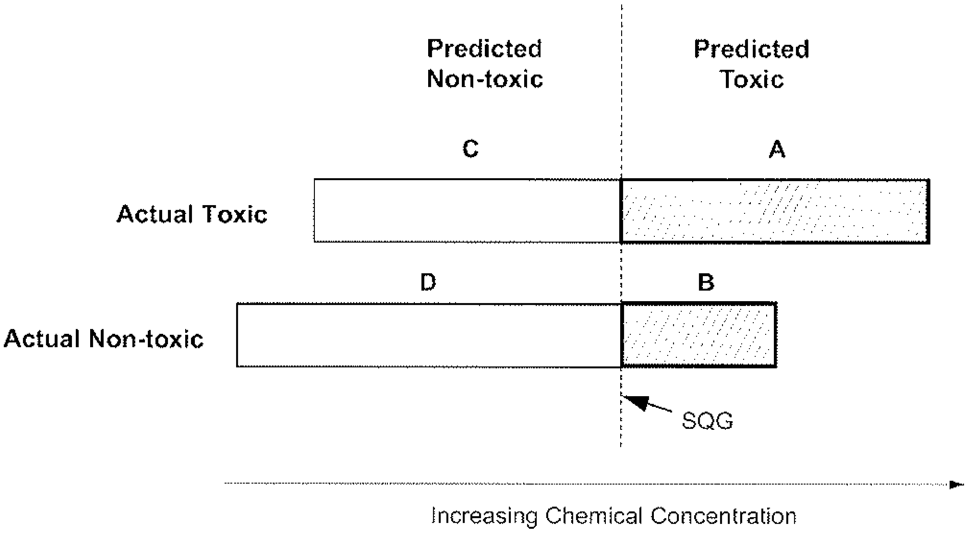
   A = toxic samples predicted as toxic
   B = non-toxic samples predicted as toxic (false positives)
   C = toxic samples predicted as non-toxic (false negatives)
   D = non-toxic samples predicted as non-toxic

These counts were used to compute the performance metrics defined in Table 7 and shown in Figure 9.

Table 7. Performance evaluation metrics defined

False Positives:	the percent of non-toxic samples incorrectly predicted to be toxic [B/(B+D)]
False Negatives:	the percent of toxic samples incorrectly predicted to be non-toxic [C/(A+C)]
Sensitivity:	the percent of toxic samples correctly predicted (100% - % false negatives) [A/(A+C)]
Toxic Efficiency	the percent of samples predicted as toxic that were actually toxic [A/(A+B)]
Non-Toxic Efficiency:	the percent of samples predicted as non-toxic that were actually non-toxic [D/(C+D)]
Predictive Reliability or Accuracy:	the percent of samples that were correctly predicted (i.e., the number of toxic samples predicted as toxic plus the number of non-toxic samples predicted as non-toxic divided by the total number of samples) [(A+D)/(A+B+C+D)]

For the majority of the SSG sets defined above (i.e., Sets #1 through #8, Table 6), a sample was predicted to be toxic if one or more individual chemical guidelines was exceeded, and non-toxic if no chemical guidelines were exceeded. The mean SQG quotient approach (Set #9 in Table 6) computes for each sample a single value that is the average of nine chemical concentrations each divided by their chemical guideline. For this approach, a sample was predicted to be toxic if the mean SQG quotient (rounded to 2 decimals) was greater than 0.10 for surface sediments or 0.50 for foundation sediments. The selection of the 0.10 threshold for surface sediments was based on the work of Fairey et al. (2001) who found that in the national databases, the incidence of toxicity was low (<5%) and average survival was high (>80%) when mean SQG quotients were below 0.10. This is consistent with the objective of using only the cleanest sediments for wetland surface material. The selection of the 0.50 threshold for foundation sediments was also based on Fairey et al.'s (2001) results as well as the response of our data set and represented an intermediate region for identifying moderately contaminated sediments. Approximately 74% of the mean SQG quotient results for samples in this database were between 0.10 and 0.50, while fewer than 10% had mean quotients exceeding 0.50 (Appendix C; Figure C-4). A threshold below 0.50 for foundation material would tend to be too conservative chemically and result in an elevated false positive rate which runs contrary to the objectives set for foundation material (Section 3.2.2), whereas a higher threshold would tend to be too restrictive and result in very few samples identified as appropriate for the wetland foundation application.



False Positives = B/(B+D)
False Negatives = C/(A+C)
Sensitivity = A/(A+C)
Toxic Efficiency = A/(A+B)
Non-Toxic Efficiency = D/(C+D)
Predictive Accuracy = (A+D)/(A+B+C+D)

Figure 9. Schematic showing calculations of performance metrics shown in Table 7

Data reported as below detection limits were included in the performance evaluations at ½ of the detection limit. All samples that had at least one chemical reported were included in all performance evaluations except the SQG-Q1 (Fairey 2001). This was done to ensure that a sample could be predicted as toxic even if it only had results for a single chemical endpoint. If the limited chemical results were not associated with a sample's toxicity, then false negative rates may be inflated over what they would be if more complete chemical results were available for that sample. This approach to sample selection may result in increased false negative rates for a set of SSGs but it provides a more complete assessment of false positives and true positives than would be achieved by excluding the samples with incomplete analyte lists. The SQG-Q1 method includes only those samples that have all nine chemicals measured so that the mean quotient is based on comparable data for all samples, per the approach used in Fairey et al. (2001).

3.2.1 Screening Guidelines for Surface Material

Wetland surface material will have direct exposure to organisms, so only the cleanest sediments should be used for this application. An SSG set useful for identifying suitable surface material should:

- Avoid predicting contaminated sediments as non-toxic (low false negatives)
- Correctly predict as many clean sediments as possible (high non-toxic efficiency)

These objectives are represented in the idealized pie chart shown in Figure 10. Results for the remaining SSG sets identified as potential surface screening numbers are shown in Figure 11 and in Table 8, and are summarized below:

- The surface SQG-Q1 predictions (i.e., mean quotient ≤ 0.10) resulted in 9% false negatives and 81% non-toxic efficiency. This guideline set predicted a total of 58 samples (with 81% accuracy) as suitable for surface material. The nine chemical endpoints included in this guideline set were chosen by the original investigators (Fairey et al. 2001) because they performed best at predicting amphipod acute toxicity and non-toxicity in national data sets. The results for this data set suggest that these nine chemical endpoints may not be the best predictors of non-toxicity under local conditions.

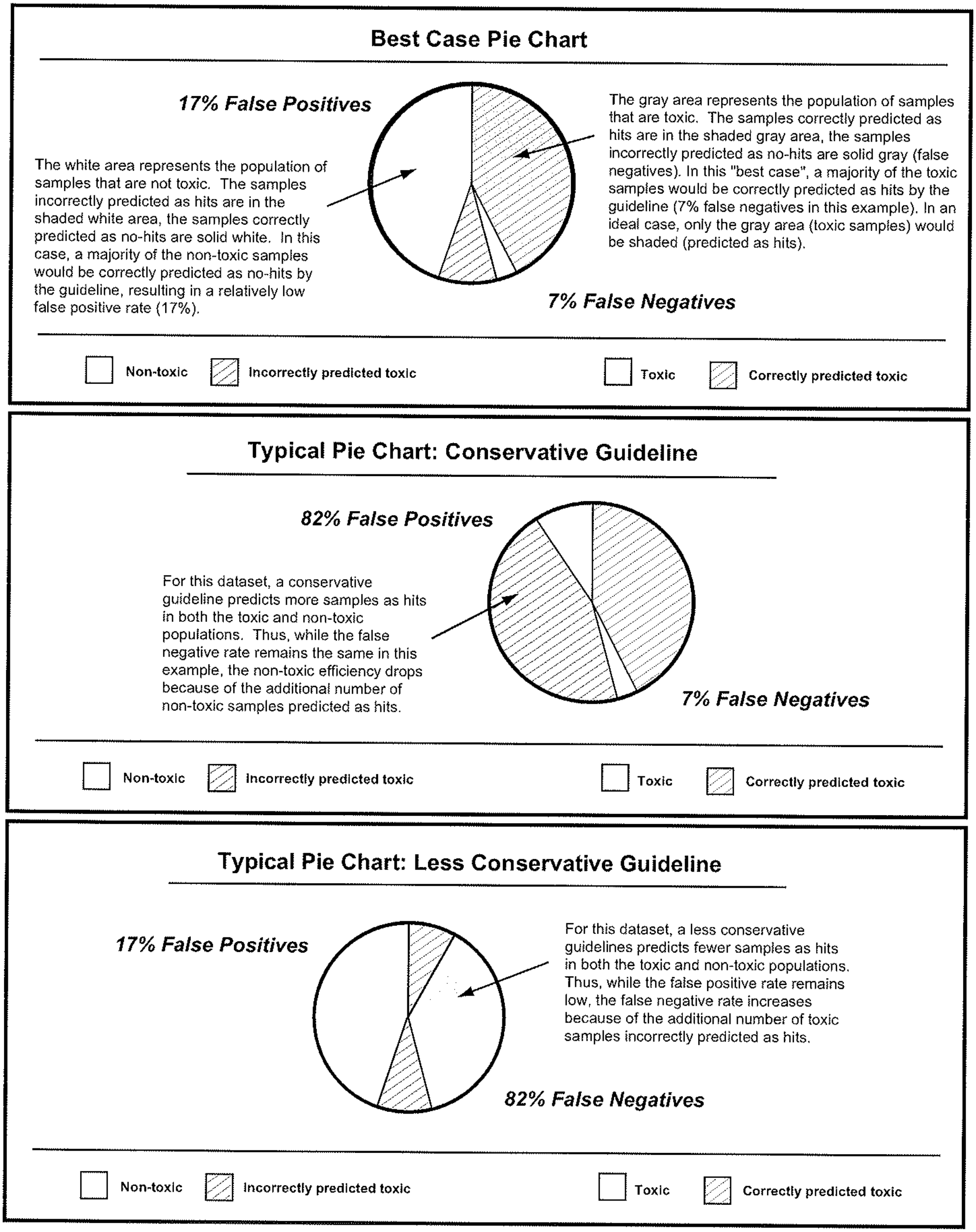
- The revised wetland surface SSG values established by the SFB-RWQCB (2000) had the best performance of all the existing SSG sets. A total of 57 samples were predicted as suitable for surface material with 84% non-toxic efficiency and 7% false negatives. These guidelines would safely identify non-toxic samples for surface use in wetland restoration. However, given the high false positive rate for these guidelines (76%), a permit applicant should always opt for bioeffects testing instead of accepting the prediction of acute toxicity based on sediment chemistry. If the sample did not exhibit adverse biological effects, then despite the elevated chemistry, the sediment would still be considered suitable for surface material (see Section 5.2).

3.2.2 Screening Guidelines for Foundation Material

No organisms would be directly exposed to wetland foundation material once it is in place, so the objective would be to make sure that a minimal amount of non-toxic sediments are classified as foundation material (these are better used as surface material). An SSG set useful for identifying suitable foundation material should:

- Avoid predicting clean sediments as toxic (low false positives)

- Correctly predict as many contaminated sediments as possible (high toxic efficiency and high sensitivity)



Best Case Pie Chart

17% False Positives

The white area represents the population of samples that are not toxic. The samples incorrectly predicted as hits are in the shaded white area, the samples correctly predicted as no-hits are solid white. In this case, a majority of the non-toxic samples would be correctly predicted as no-hits by the guideline, resulting in a relatively low false positive rate (17%).

The gray area represents the population of samples that are toxic. The samples correctly predicted as hits are in the shaded gray area, the samples incorrectly predicted as no-hits are solid gray (false negatives). In this "best case", a majority of the toxic samples would be correctly predicted as hits by the guideline (7% false negatives in this example). In an ideal case, only the gray area (toxic samples) would be shaded (predicted as hits).

7% False Negatives

Typical Pie Chart: Conservative Guideline

82% False Positives

For this dataset, a conservative guideline predicts more samples as hits in both the toxic and non-toxic populations. Thus, while the false negative rate remains the same in this example, the non-toxic efficiency drops because of the additional number of non-toxic samples predicted as hits.

7% False Negatives

Typical Pie Chart: Less Conservative Guideline

17% False Positives

For this dataset, a less conservative guidelines predicts fewer samples as hits in both the toxic and non-toxic populations. Thus, while the false positive rate remains low, the false negative rate increases because of the additional number of toxic samples incorrectly predicted as hits.

82% False Negatives

Figure 10. Graphical Guideline Performance: Understanding the Pie Chart Displays

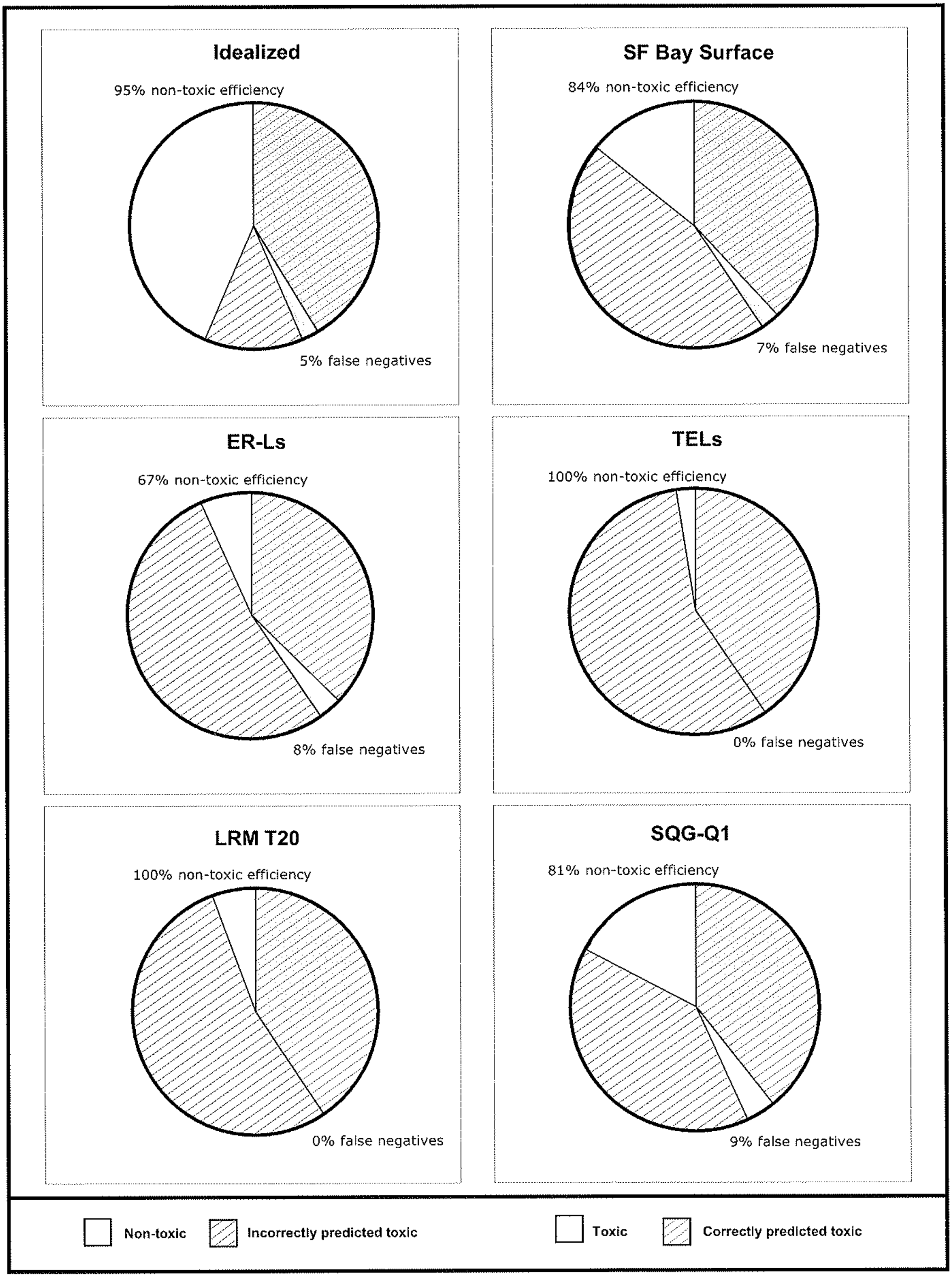


Figure 11. Performance evaluation results for potential surface screening numbers

Table 8. Performance metrics for potential wetland surface and foundation sediment screening guidelines

		Sample Counts						Performance Metrics (%)					
SSG Values	Prediction Method[1]	Hits	No Hits	Hits Pred Hits	NoHits Pred Hits	Hits Pred NoHits	NoHits Pred NoHits	Sensitivity	Toxic Efficiency	Non-Toxic Efficiency	Predictive Accuracy	False Positives	False Negatives
		Formulas:		*A*	*B*	*C*	*D*	*=A/(A+C)*	*=A/(A+B)*	*=D/(C+D)*	*=(A+D)/ (A+B+C+D)*	*=B/(B+D)*	*=C/(A+C)*
Pooled Amphipod Toxicity													
Surface Material Values													
LRM T20	Single Exceedance	137	199	137	180	0	19	100	43	100	46	90	0
ER-L	Single Exceedance	137	200	126	178	11	22	92	41	67	44	89	8
TEL	Single Exceedance	137	200	137	192	0	8	100	42	100	43	96	0
Existing Surface	Single Exceedance	137	200	128	152	9	48	93	46	84	52	76	7
SQG-Q1	Mean Quotient > 0.10	119	155	108	108	11	47	91	50	81	57	70	9
Floating Percentile[2]	Single Exceedance	137	199	128	143	9	56	93	47	86	55	72	7
Foundation Material Values													
LRM T40	Single Exceedance	137	199	118	128	19	71	86	48	79	56	64	14
LRM Pmax 40%	Single Exceedance	137	199	92	82	45	117	67	53	72	62	41	33
PEL	Single Exceedance	137	200	80	105	57	95	58	43	63	52	53	42
Existing Foundation	Single Exceedance	137	200	29	36	108	164	21	45	60	57	18	79
SQG-Q1	Mean Quotient > 0.50	119	155	18	5	101	150	15	78	60	61	3	85
Floating Percentile	Single Exceedance	137	199	33	26	104	173	24	56	62	61	13	76

[1] Prediction methods for a "hit" or toxic sample are:
Single Exceedance = one or more individual chemical guidelines exceeded
Mean Quotient (mean of concentrations divided by guidelines, rounded to 2 decimals) exceeds specified threshold

[2] Floating Percentile values are the site-specific values generated from this dataset. These are discussed and presented in Section 5.1

Notes:

All results use the 10th percentile tolerance limit for A. abdita and the 20th percentile tolerance limit for E. estuarius (see Section 4.1; also Appendix C, Section 3.2)
For individual compounds, all results use ½ the detection limit for data below detection.
All results except SQG-Q1 compute totals excluding data below detection limits; SQG-Q1 used methods to match Fairey et al. (2001) (see text Appendix C, Section 2.1).

A = Hits (toxic samples) predicted as Hits (exceeding guidelines by the specified prediction method)
B = No Hits (non-toxic samples) predicted as Hits
C = Hits predicted as No Hits
D = No Hits predicted as No Hits

These objectives are represented in the idealized pie chart shown in Figure 10. Results for the remaining SSG sets identified as potential foundation screening numbers (Table 6) are also shown in Figure 11 and in Table 8, and are summarized below.

- The SSG sets based on the logistic regression models LRM T40 (T40 = concentration associated with a 40% probability of acute toxicity) and LRM P_{max} 40% (multi-chemical model associated with a 40% probability of toxicity adjusted for overestimation) performed poorly with fairly high false positives (64% and 41%, respectively), moderate toxic efficiencies (48% and 53%, respectively), and moderate to good sensitivities (86% and 67%, respectively). For foundation screening, these guidelines are clearly too low on the LRM dose-response curves.

- The PEL values also performed poorly with 53% false positives, 43% toxic efficiency, and 58% sensitivity.

- The foundation SQG-Q1 predictions (i.e., mean quotient > 0.50) performed the best of all potential foundation guidelines from a rate perspective with 3% false positives and 78% toxic efficiency (Figure 12). However, sensitivity was a low 15%, and only 23 out of 337 samples were predicted as suitable for foundation material. Most of the toxic samples in the database had concentrations below these guidelines, which means that the observed acute toxicity was not likely due to the compounds on the SQG-Q1 list at those threshold levels.

- The set of existing foundation SSG values had poor reliability results with 18% false positives, 45% toxic efficiency, and only 21% sensitivity (Figure 12). A total of 65 samples (with 45% accuracy) were predicted as suitable for foundation material using these guidelines.

The performance of all of the existing SSG sets suggests that the amphipod toxic responses cannot be explained by the action of individual chemicals alone; none of the existing guidelines were good predictors of amphipod acute toxicity, and none of them were able to simultaneously achieve a low false positive rate and high sensitivity. This apparent lack of distinction in chemical concentrations between the toxic and non-toxic samples is graphically illustrated by chemical distribution plots (Figure 13). These plots allow a comparison of the distribution of concentrations in the toxic and non-toxic samples for individual chemical endpoints. These plots are an over-simplification of the relationship between acute toxicity and concentrations for individual chemicals in the chemical mixtures. For any given chemical, acute toxicity may be observed in samples with low concentrations of that particular chemical but be caused by elevated concentrations of another chemical in the same sediment. These plots do provide some indication of which chemicals may be driving toxicity. A chemical driving acute toxicity would have many toxic samples found at concentrations above the highest non-toxic sample concentration.

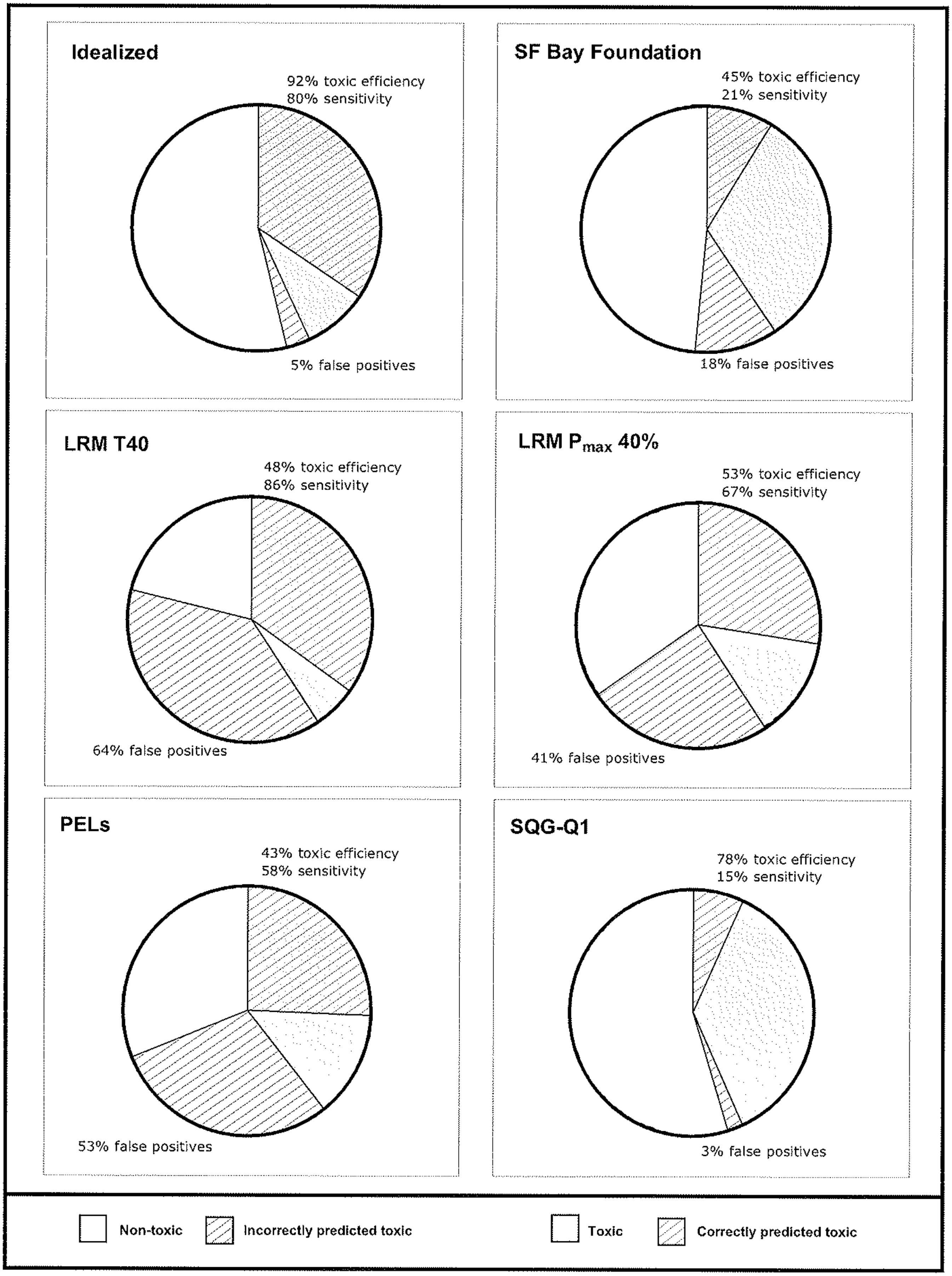
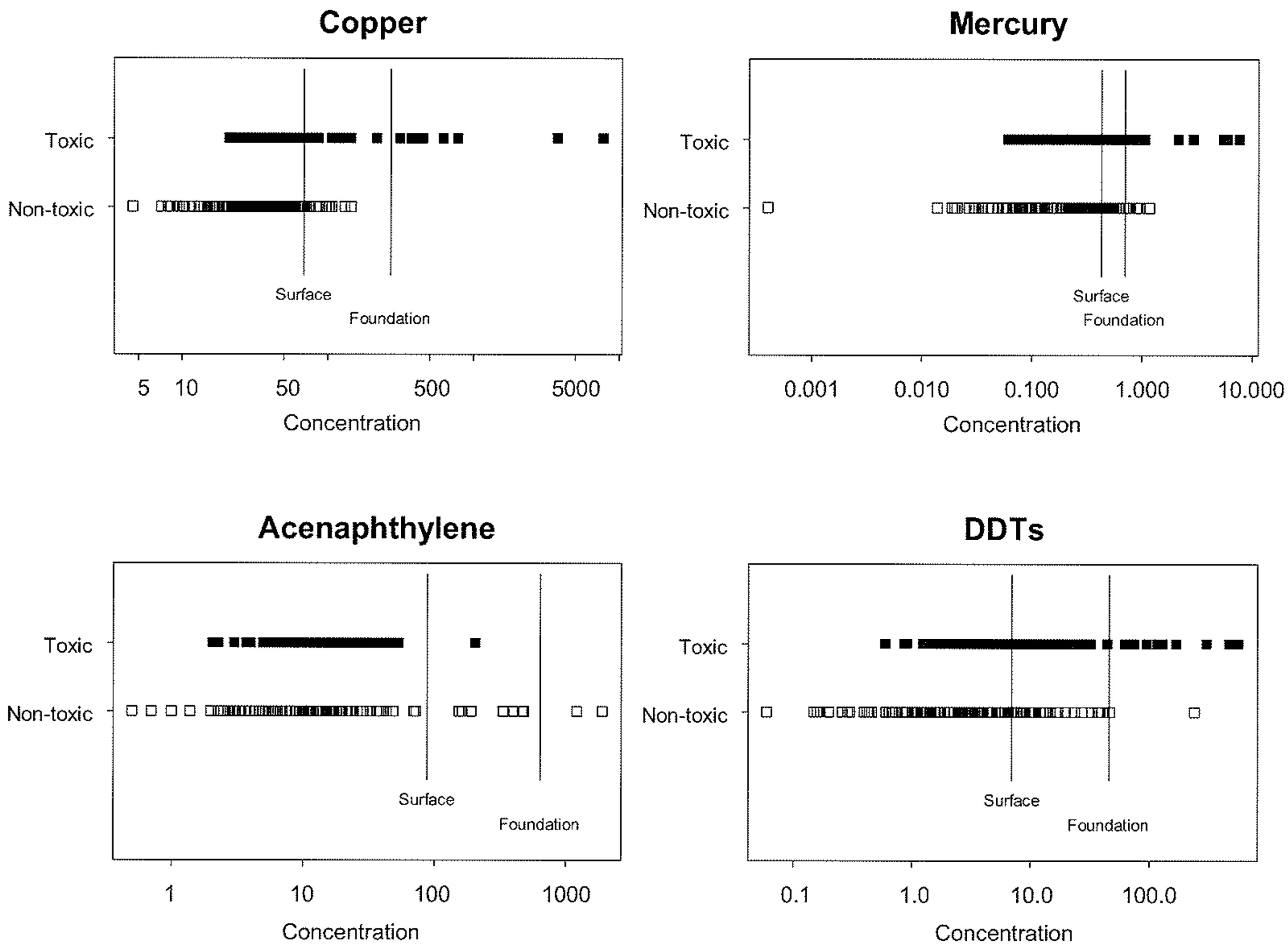


Figure 12. Performance evaluation results for potential foundation screening numbers



Note: Toxicity determined by pooled amphipod endpoint using final toxicity thresholds for *A. abdita* and *E. esturarius*; data below detection limits are excluded.

Figure 13. Distribution plots for selected chemical endpoints

Distribution plots for several chemicals are shown in Figure 13 to illustrate the typical patterns observed in this data set (plots for all chemical endpoints in the database are included in Appendix C, Figure C-1). All of the chemical endpoints showed substantial overlap in concentration ranges between the toxic and non-toxic samples. Copper and mercury were the only two endpoints that were most clearly associated with acute toxicity at the upper ranges of their concentrations. For the majority of chemical endpoints, the overlap between the two subsets of samples is nearly complete, and in several cases the sediment concentrations in the non-toxic samples exceed the concentrations in the toxic samples (e.g., acenaphthylene).

There are several possible explanations for the observed concentration overlap for so many chemical endpoints:

1. The concentrations do not consistently extend into the acutely toxic range (the range in values is not great enough; the database needs more representative samples from areas with higher contamination).

2. Acute toxicity is caused primarily by chemicals that are not being measured in the standard suite of analytes used for routine bulk sediment analyses.

3. Acute toxicity is caused by synergistic or antagonistic effects of measured chemical concentrations, other physical factors (e.g., grain size, TOC), and possibly unknown or unmeasured factors.

4. The measured chemicals are indeed responsible for adverse biological effects, but the test organisms used are imprecise and unreliable indicators of acute toxicity.

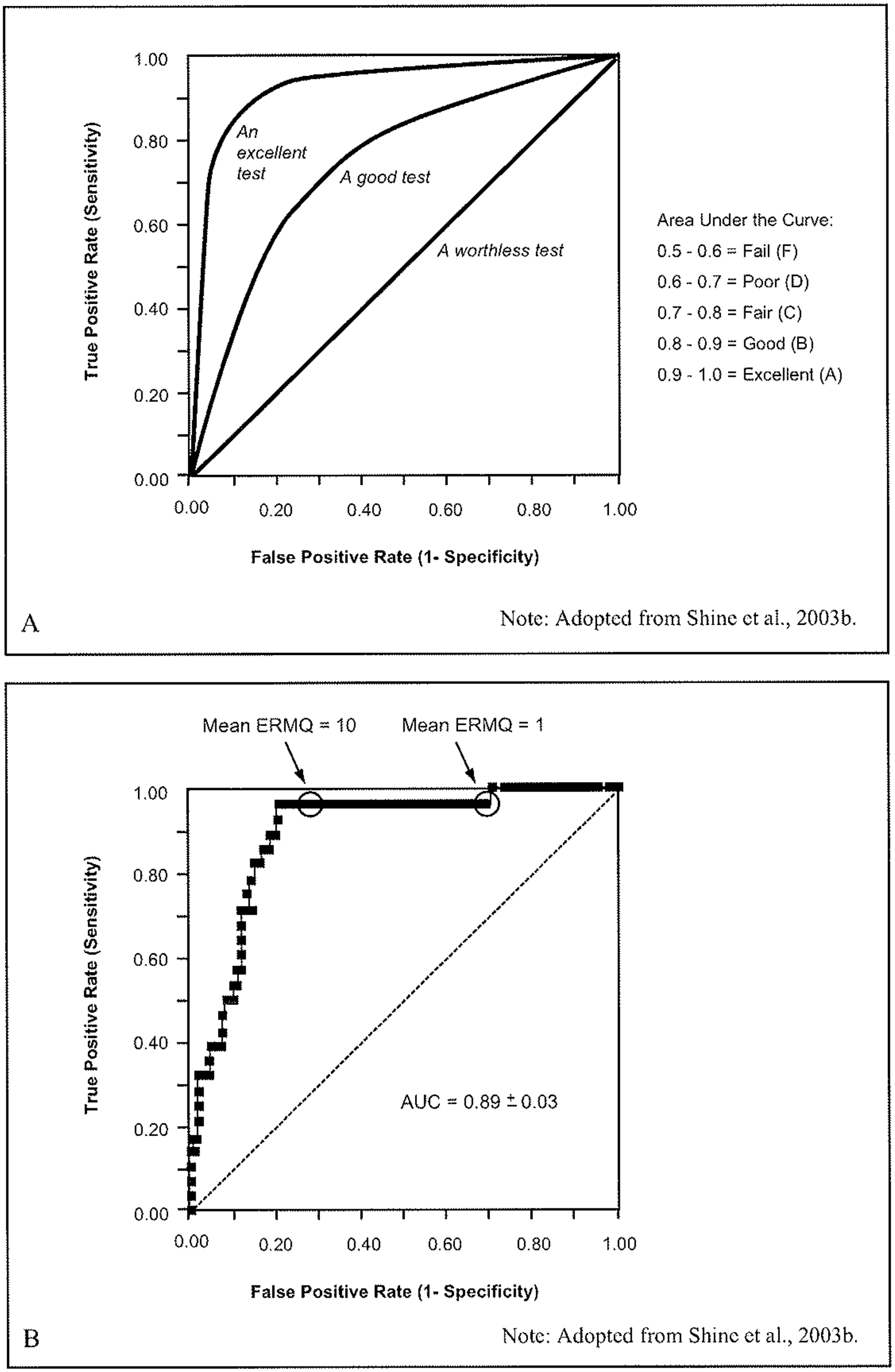
Whatever the reason, the substantial overlap in concentration distributions observed in the majority of the distribution plots illustrates why the application of SSG values to this particular database will always result in high error rates of one type or another.

3.3 Receiver Operating Characteristic Curves

3.3.1 Method Description

We also investigated methods for identifying the chemicals or group of chemicals most predictive of toxicity. A method commonly used in the biomedical field for assessing the discriminatory power of diagnostic tests called Receiver Operating Characteristics (ROC) was adapted for assessment of SSGs (Shine et al., 2003b).

ROC curves were used by Shine et al. (2003b) to evaluate SSGs for metals by revealing compromises in sensitivity (the ability to correctly classify a toxic sample as toxic, or true positive) and specificity (the ability to correctly classify a non-toxic sample as non-toxic, or true negative) associated with a given chemical concentration or threshold. The resulting shape of the curve is an indicator of how well that guideline distinguishes between the false positive and true positive endpoints (Figure 14A).



Note: Adopted from Shine et al., 2003b.

Note: Adopted from Shine et al., 2003b.

Figure 14. Idealized plots of ROC curves (A) with the interpretation of the AUC and (B) showing optimization of a guideline

To understand the idealized curve in Figure 14A, it is important to understand the components underlying the data that generate the curve. First, only samples that are *predicted to be toxic* by that guideline are evaluated. Therefore, the test is useful in evaluating SSGs for their ability to predict toxicity, excluding analysis of non-toxic response. Further, of those samples predicted to be toxic by a given guideline, some proportion is actually toxic (true positive rate), and some proportion is not (false positive rate).

Each point along the curve in Figure 14A represents a potential SSG value, increasing in concentration from right to left. By definition, at the far right of the graph and the minimum potential SSG (defined by the lowest value in the database), 100% (normalized to 1 in Figure 14A) of the samples will be plotted as they will all be predicted to be toxic (e.g., greater than the SSG). Of those, all the actual toxic samples will have been correctly predicted (100% true positive rate), and all of the non-toxic samples will be incorrectly predicted (100% false positive rate). At the other end of the curve, at the maximum SSG value (defined by the highest value in the database), there will be no samples that are predicted to be toxic, so both the false positive and true positive rates are zero, by definition. The shape of the curve between these two endpoints defines how these two rates vary between the potential guidelines. An "excellent," or predictive, chemical or quotient results when the false positive rate drops more rapidly than the true positive as the guideline increases.

Shine et al. (2003a) evaluated the shape of the ROC curve (Figure 14A), using the calculation of the area under the curve (AUC). They suggested a scale for evaluation of the final AUC values: 0.5-0.6 fail; 0.6-0.7 poor; 0.7-0.8 fair; 0.8-0.9 good; 0.9-1.0 excellent.

The ROC curves were created by plotting sensitivity (true positives) against 1-specificity (false positives) as described by Shine et al. (2003 a,b). The area under the ROC curve was calculated using the trapezoid rule on the empirical curve. The AUC can range from 0.0 to 1.0; the closer the value is to 1.0, the more effective the test – that is, the better the chemical or threshold minimizes both false positives and false negatives. Any value <0.5 indicates that the concentrations in the non-toxic samples exceed the concentrations in the toxic samples.

We conducted ROC analyses using the pooled amphipod endpoint (tolerance limit based on the 20th percentile for *E. estuarius* and 10th percentile for *A. abdita*), excluding data below detection. Chemical-specific ROC curves were generated only for chemicals with at least 100 samples, because a curve formed by too few samples can cause anomalous AUC values. The data for the curves were generated by evaluating every existing value of a selected chemical in the database as a potential guideline, then summarizing the performance metrics (in terms of false and true positives) for each sample. Because the analysis was very data intensive, several programs, queries, and graph macros were written in support of the ROC analyses and output curves (Section 2.0).

In addition to finding toxicity drivers, we hoped to use this method to optimize the selected guideline. For example, a potential guideline value theoretically can be increased (from the ERMQ of 1 to 10 in Figure 14B) without any compromise in the true positive rate. Ideally, then, the guideline could be maximized at the break in the curve where the true positive rate begins to decrease. This method could be used for a single chemical guideline, or a quotient type guideline approach.

3.3.2 ROC Results

We plotted all the ROC curves for chemicals with at least 100 samples (Appendix C, Figure C-6) and calculated the areas under the curves for those chemicals (Table 9). A typical example of a chemical-specific ROC curve for our database is shown for total PAHs (dry weight) in Figure 15. The area under the curve for this chemical was 0.72, one of the highest values for the project database, although only 'fair' using Shine et al.'s (2003a) ranking. In fact, no chemical with at least 100 results (too few results tended to highly skew the shape of the curve) resulted in areas under the ROC curve of >0.75, and of those chemicals with AUC values of >0.7, all were PAHs (Table 9). The AUC index is used as an easy reference for the shape of the curve. An ideal curve that rises quickly to the asymptote of 1.0 (as in Figure 14B) will have a higher AUC than one that rises gently and steadily (see Figure C-6 for ROCs from this data set).

Table 9. Results of area under the curve analyses for chemicals with at least 100 samples

Chemical Name	Area Under the Curve (AUC)	Number of Samples
1-Methylphenanthrene	**0.74**	**127**
2-Methylnaphthalene	**0.74**	**101**
4,4'-DDD	0.68	177
4,4'-DDE	0.67	179
Acenaphthene	0.61	142
Acenaphthylene	0.58	146
Aluminum	0.54	154
Anthracene	0.67	179
Arsenic	0.48	218
Benz(a)anthracene	**0.73**	**244**
Benzo(a)pyrene	**0.73**	**254**
Benzo(b)fluoranthene	**0.72**	**264**
Benzo(e)pyrene	**0.75**	**193**
Benzo(g,h,i)perylene	**0.71**	**250**
Benzo(k)fluoranthene	**0.73**	**230**
Cadmium	0.67	199
Chromium	0.60	240
Chrysene	**0.72**	**252**
Clay	0.66	145
Copper	0.58	245
DDTs	0.70	208
Dibenz(a,h)anthracene	0.66	174
Fines	0.58	228
Fluoranthene	**0.72**	**251**
Fluorene	0.61	148
HPAH	**0.73**	**301**
Indeno(1,2,3-c,d)pyrene	**0.72**	**259**
Iron	0.56	144
Lead	0.60	265
LPAH	0.67	279
Manganese	0.51	127
Mercury	0.57	216
Naphthalene	**0.71**	**178**
Nickel	0.52	238
PAHs	**0.72**	**307**
PCBs	0.67	195
Perylene	0.70	164
Phenanthrene	**0.71**	**239**
Pyrene	**0.74**	**269**
Sand	0.31	122
Selenium	0.51	142
Silt	0.60	107
Silver	0.61	216
Solids	0.39	143
TOC	0.60	156
Total PAHs (molar)	0.69	311
Zinc	0.63	278

Chemicals in bold and highlighted have AUC values of >0.7.

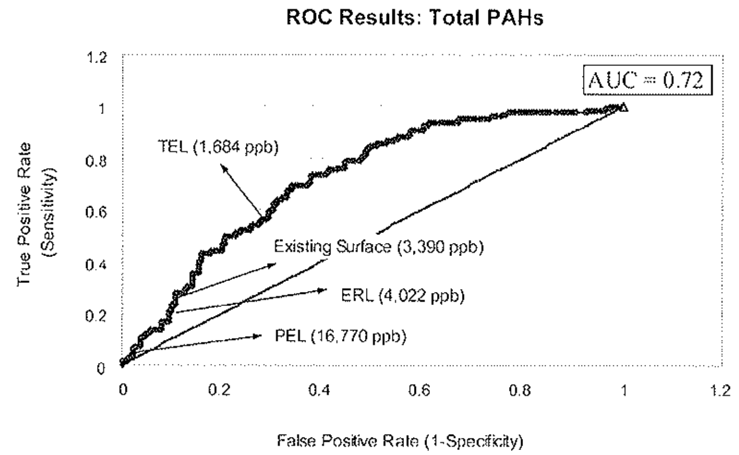


Figure 15. ROC curve for total PAHs (dry weight), showing the curve, the 1:1 line, the area under the curve (AUC) result, and the location of existing guideline values along the curve

Our intent to use the ROC curves to identify optimal guideline values for this dataset was unsuccessful due to the shape of the ROC curves. The gentle, flat rise of the curves indicated there was no region where a sharp increase in true positives (sensitivity) can be gained with only a minor loss in specificity, except at very low concentrations (i.e., at the right side of the curve). Essentially, the tradeoff between sensitivity and specificity was constant, and an ideal break point (for example see Figure 14B) was not found in these data.

The ROC approach appears to be a useful method both to quantify and graphically illustrate the relative sensitivity and specificity in a data set and may prove to be a useful tool to develop better SSGs in a given database. However, consistent with the other data evaluation approaches taken, the results suggest that the project data lack a sufficient relationship with chemical concentrations to be able to identify any chemicals or groups of chemicals that accurately predict toxic or non-toxic responses.

4.0 SITE-SPECIFIC GUIDELINE DEVELOPMENT

From the preceding results, it was apparent that existing SSGs were not reliable for predicting toxic and non-toxic samples in the San Francisco database. Some of the existing SSGs had fairly good rates for their intended purpose (e.g., low false negatives and high non-toxic efficiency for surface material) but the number of samples predicted for either surface or foundation materials was generally quite low. In addition, when false negatives were low (a surface guideline objective), then false positives were very high (meaning many of the non-toxic sediments were missed by the surface screening guidelines). False negatives and false positives will always be inversely related within a given data set; this inverse relationship becomes more pronounced with data sets that have substantial overlap in concentration ranges, such as was the case in the San Francisco Bay data set assembled for this project. Consequently, we expected these data to show this costly trade-off between false negatives and false positives. Simultaneous optimization of these error rates can be achieved through site-specific SSG development using the Floating Percentile method. Because this method considers all percentiles for each individual chemical, it provides the best combination of performance metrics that is possible for these data (i.e., for a fixed false negative rate, one can determine the lowest possible false positive rate).

4.1 Materials and Methods

The Floating Percentile method uses different percentiles for each chemical in the complete synoptic data set (i.e., all toxic and non-toxic stations). The basic concept behind the Floating Percentile method is to select a fixed percentile of the data that provides a low false negative rate, then adjust individual chemical values upward until false positive rates are optimized (decreased to their lowest possible level) while retaining the same level of false negatives. Once each chemical has been individually adjusted upward to its threshold, the false positives will have been significantly reduced while retaining the same low false negative rate. In this manner, optimized criteria sets can be developed for a number of different target false negative rates, allowing the trade-offs between false negatives and false positives to be evaluated, and a final set of SSGs to be selected.

Complete details on the data preparation and computation steps involved in the derivation of Floating Percentile guidelines are presented in Appendix D.

The data used in this Floating Percentile guideline development process is a subset of the data used in the performance evaluations (see Section 3.2, Reliability Results for Existing SSGs). For the previous performance evaluations, the decision was made to include samples even if they had only one chemical endpoint and to include data below detection limits at ½ the detection limit. That approach allows the possibility that acute toxicity could be predicted in a sample with only one chemical reported; but it comes at the potential cost of increased false negative rates if unmeasured chemicals were elevated and associated with sample toxicity. In the Floating Percentile process, individual chemical thresholds are identified based on an optimization of the performance metrics computed from consideration of the complete chemical mixture for each sample in the data set. Reduced analyte lists for samples and uncertainty in the actual concentrations represented by data below detection limits will adversely affect performance metrics and confound the location of toxicity thresholds in the data set. Consequently, the Floating Percentile guideline development process excluded data below detection limits and included only those samples that had results reported for at least one metal and one PAH (184 non-toxic samples and 128 toxic samples). The data set used in the Floating Percentile guideline development process is referred to as the "restricted" data set because of the exclusion of data below detection and the restriction of samples with at least one metal and one PAH. The "complete" data set was used in all other performance evaluations.

Sums of BHCs, chlordanes, PCBs, PAHs, and DDTs were calculated using zero for data below detection limits. Several exploratory runs of the Floating Percentile method were conducted to provide a comparison of various techniques for working with the data set (results are summarized in Appendix C, Section 4.4). Observations from these exploratory runs led to the following decisions regarding specific chemical endpoints in the final Floating Percentile calculations:

- The molar sum of PAHs was used in place of individual PAHs to reflect their additive toxicity based on a narcosis toxicity model. The narcosis model is a general model of toxicity to aquatic receptors based on the presence of organic chemicals in tissues and their disruption of basic cellular functions (Connell and Markwell, 1992; Veith et al., 1983; DiToro et al., 2000). This mode of toxicity occurs in all species and is dependent only on the total molar concentration of chemicals partitioned into lipid tissues. (See page D-3 in Appendix D for more details on the narcosis toxicity model.) Specifically, recent research suggests that mixtures of PAHs may have greater toxic effects than individual PAH compounds to amphipods, and that there may be a total molar concentration threshold related to toxicity to the amphipod *Diporeia* (Landrum et al. 2003). The molar sum reflects the relative presence of higher versus lower molecular weight PAHs, and therefore the potential increased toxicity. The molar sum can be calculated readily from the dry weight concentration normalized to the molar weight of the PAH compound (Appendix C, Table C-1).

- Dibenzothiophene and selenium were dropped from the SSG analyte list after a sensitivity analysis determined that reliability improved when they were excluded. (Note: sensitivity analyses indicated that reliability declined when each of the remaining chemical endpoints associated with a large number of errors was excluded from the SSG analyte list, so those analytes were not excluded from the final SSG analyte list.)

The goal of this effort was to provide more reliable SSGs, if possible, than the existing surface and foundation SSGs for wetland reuse (SFB-RWQCB, 2000) as well as other available sediment screening guideline sets. Emphasis for the suggested surface SSGs was placed on low false negatives and high non-toxic efficiency, while emphasis for the suggested foundation SSGs was placed on low false positives, high sensitivity, and high toxic efficiency. The most useful screening guidelines will meet these goals while also optimizing the other measures of reliability to provide better overall discrimination between toxic and non-toxic sediments.

4.2 Results and Recommended Numbers

A Floating Percentile evaluation was conducted to derive SSGs for a full range of target false negative and false positive levels (from 0% false negatives to 0% false positives), as well as the four other reliability measures. The reliability metrics and the optimized chemical thresholds for 15 analytes associated with each target false negative rate (i.e., 19 discrete values ranging from 0% to 89% false negatives) are provided in Table 10. These data are illustrated in Figure 16, showing the relationships between all six performance metrics for the optimized reliability results.

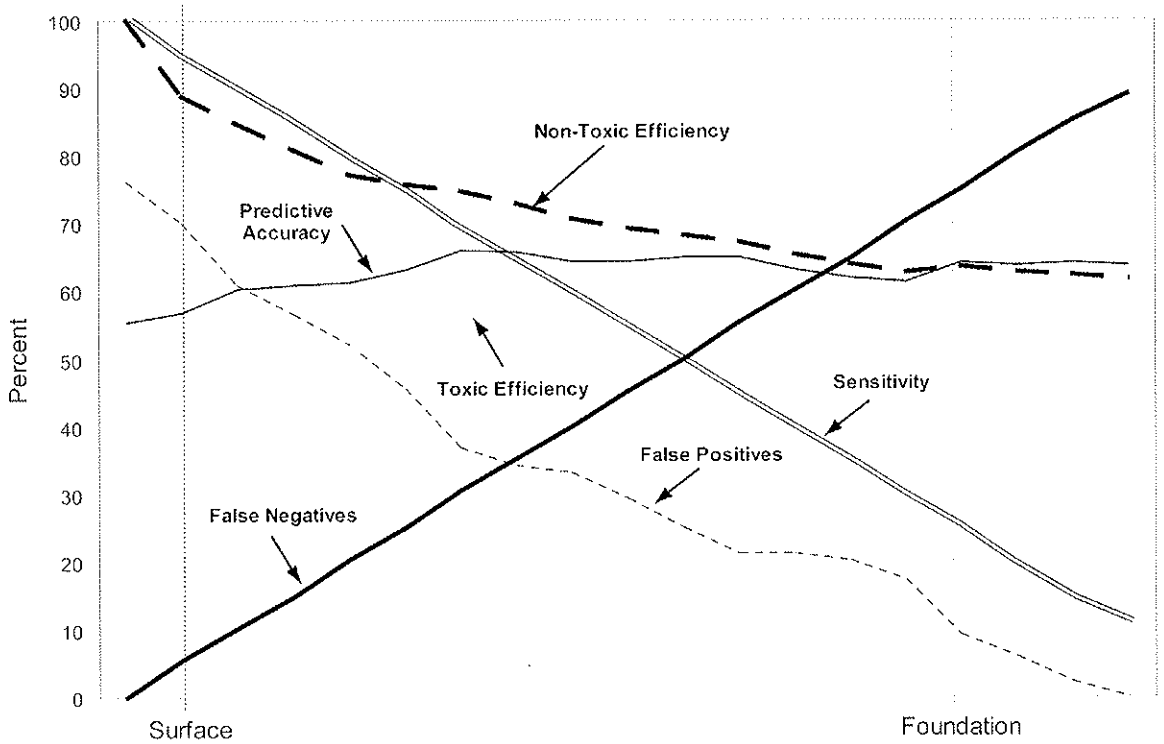
Table 10. Full range of optimized reliability results from final Floating Percentile calculations

							Metals (ppm, DW)					
Row	% False Negatives	% False Positives	% Sensitivity	% Toxic Efficiency	% Non-Toxic Efficiency	% Reliability	Arsenic	Cadmium	Chromium	Copper	Lead	Mercury
1	0	76	100	48	100	55	40	0.25	94	50	200	1.18
2	5	70	95	48	89	57	40	0.25	119	50	200	1.18
3	10	61	90	51	85	60	40	0.28	129	56	200	1.18
4	15	57	85	51	81	61	40	0.28	150	56	200	1.18
5	20	52	80	52	77	61	40	0.31	150	56	200	1.18
6	25	46	75	53	76	63	40	0.34	150	56	200	1.18
7	30	37	70	57	75	66	40	0.34	150	62	200	1.18
8	35	34	65	57	73	66	40	0.34	170	62	200	1.18
9	40	33	60	56	71	64	40	0.35	170	62	200	1.18
10	45	29	55	56	69	64	40	0.38	170	62	200	1.18
11	50	25	50	58	68	65	40	0.43	170	67	200	1.18
12	55	21	45	59	67	65	40	0.45	195	67	200	1.18
13	60	21	40	57	65	63	40	0.46	195	67	200	1.18
14	65	20	35	55	64	62	40	0.46	195	67	200	1.18
15	70	17	30	54	63	61	40	0.58	195	67	200	1.18
16	75	9	25	65	63	64	40	0.62	320	150	200	1.18
17	80	6	20	69	63	64	40	0.80	320	150	200	1.18
18	85	2	15	83	62	64	40	0.90	320	150	200	1.18
19	89	0	11	100	62	64	40	4.00	320	150	200	1.18

Note: Shaded rows highlight the FP SSG sets recommended for surface and foundation screening guidelines.

Table 10, continued

Metals (ppm, DW)			PAHs (DW)	Chlorinated Organic Compounds (ppb, DW)	Pesticides and PCBs (ppb, DW)			
Nickel	Silver	Zinc	Total PAHs (molar)	Hexachloro-benzene	Chlordane	Total BHCs	Total DDTs	Total PCBs
230	0.28	1200	6.3	60	69.2	2.00	250	600
230	0.28	1200	6.3	60	69.2	2.00	250	600
230	2.00	1200	6.3	60	69.2	2.00	250	600
230	2.00	1200	7.1	60	69.2	2.00	250	600
230	2.00	1200	7.7	60	69.2	2.00	250	600
230	2.00	1200	9.5	60	69.2	2.00	250	600
230	2.00	1200	10.6	60	69.2	2.00	250	600
230	2.00	1200	11.0	60	69.2	2.00	250	600
230	2.00	1200	12.0	60	69.2	2.00	250	600
230	2.00	1200	14.0	60	69.2	2.00	250	600
230	2.00	1200	15.0	60	69.2	2.00	250	600
230	2.00	1200	16.0	60	69.2	2.00	250	600
230	2.00	1200	17.0	60	69.2	2.00	250	600
230	2.00	1200	21.0	60	69.2	2.00	250	600
230	2.00	1200	23.0	60	69.2	2.00	250	600
230	2.00	1200	32.0	60	69.2	2.00	250	600
230	2.00	1200	45.0	60	69.2	2.00	250	600
230	2.00	1200	250.0	60	69.2	2.00	250	600
230	2.00	1200	250.0	60	69.2	2.00	250	600



Note: The X-axis is categorical with each point corresponding to a row in Table 10 showing a set of Floating Percentile guidelines and the corresponding reliability results.

Figure 16. Reliability measures for the San Francisco Bay pooled amphipod data set

As stated earlier, the objectives for the suggested surface guidelines were low false negatives and high non-toxic efficiency; objectives for the suggested foundation guidelines were low false positives, high sensitivity, and high toxic-efficiency. Given these objectives, the two sets of floating percentile SSGs with reliability results summarized in Table 11 are recommended for use. The chemical concentrations for the final list of target analytes for these Floating Percentile optimized SSG sets are shown in Table 12. Table 11 shows the reliability results for the Floating Percentile optimized SSGs in comparison to the existing wetland screening guidelines for surface and foundation material in San Francisco Bay.

As shown in Table 11, the Floating Percentile process provided a significant improvement in false negatives and non-toxic efficiency for the suggested surface SSGs. Modest improvements were found for sensitivity, toxic efficiency, and overall reliability. These screening guidelines predicted a total of 61 samples from the restricted data set, or 65 samples from the complete data set, as eligible for reuse as wetland surface material.

A significant improvement was also obtained for the suggested foundation SSGs for the metrics of primary consideration for this application. The false positive rate decreased from 18% to 9%, and the toxic efficiency substantially increased from 45% to 65%. Modest improvements were found for false negatives, non-toxic efficiency, and overall reliability. These screening guidelines predicted a total of 48 samples from the restricted data or 59 samples from the complete data set, as eligible for reuse as wetland foundation material. Additional possibilities for foundation sediments would include SSG sets with 5% or 0% false positives, which can be reviewed in Table 10; however these screening guideline sets would result in even fewer samples eligible for wetland foundation.

Table 11. Reliability results for Floating Percentile SSGs[1]

SSGs	% False Negatives	% False Positives	% Sensitivity	% Toxic Efficiency	% Non-Toxic Efficiency	% Reliability
Surface Screening Guidelines						
Existing SF Bay	16	67	84	46	76	54
Floating Percentile	5	70	95	48	89	57
Foundation Screening Guidelines						
Existing SF Bay	78	18	22	45	61	58
Floating Percentile	75	9	25	65	63	64

[1] These results based on the restricted dataset used to derive the Floating Percentile guidelines. This restricted data set consisted of 312 samples (184 non-toxic and 128 toxic) for the pooled amphipod endpoint; each sample had at least one metal and one PAH reported; all data below detection limits were excluded. Reliability results based on all 336 samples with one or more chemical (complete data set matching reliability results reported in Section 4.2) are shown in Table 8 and Figure 15.

The reliability results for the existing San Francisco Bay wetland screening guidelines and the final suggested Floating Percentile screening guidelines for surface and foundation are shown in Figure 17. These reliability results match those shown in Table 8 (i.e., they are based on the complete synoptic data set of 336 samples).

Figures 18 and 19 show the percent change in threshold values between the site-specific Floating Percentile values and the historical San Francisco Bay wetland screening guidelines (SFB-RWQCB, 1992, 2000). Some of the threshold numbers change dramatically. When the threshold numbers change dramatically with only a marginal, but positive effect on errors, this means that those chemicals are not strongly associated with acute toxicity in this data set. While bioaccumulation concerns will likely result in lower trigger values for some of the persistent organic pollutants (e.g., DDTs and PCBs), these suggested Floating Percentile guidelines represent the acute toxicity thresholds in this data set.

Some of the suggested guidelines are identical for surface and foundation applications. In fact, for several compounds (e.g., arsenic and total BHCs), the suggested guideline concentrations did not change across the entire range of errors (Table 10). This occurs for the compounds for which the toxicity threshold is fairly well-defined in the data set. All toxicity thresholds are conditional on the concentrations for the rest of the chemicals in the mixture.

Table 12. Final optimized Floating Percentile SSGs and historical San Francisco Bay wetland screening criteria

	Surface Values			Foundation Values		
Chemical Name	2003 Floating Percentile	2000 SF Bay[1]	1992 SF Bay[2]	2003 Floating Percentile	2000 SF Bay[1]	1992 SF Bay[2]
Metals (ppm, DW)						
Arsenic	40.0	15.3	33.0	40.0	70.0	85.0
Cadmium	0.250	0.330	5.00	0.620	9.60	9.00
Chromium	119	112	220	320	370	300
Copper	50.0	68.1	90.0	150	270	390
Lead	200	43.2	50.0	200	218	110
Mercury	1.18	0.430	0.350	1.18	0.700	1.30
Nickel	230	112	140	230	120	200
Silver	0.280	0.580	1.00	2.00	3.70	2.20
Zinc	1,200	158	160	1,200	410	270
Total PAHs (ppb, DW)	6.3 (molar sum)	3,390	4,000	32 (molar sum)	44,792	35,000
Chlorinated organic compounds (ppb, DW)						
Hexachlorobenzene	60	0.485	--	60	--	--
Pesticides and PCBs (ppb, DW)						
Total DDTs	250	7.0	3.0	250	46.1	100
Chlordane	69.2	2.3	--	69.2	4.8	--
Total BHCs	2.0	--	--	2.0	--	--
PCBs	600	22.7	50	600	180	400

[1] SFB-RWQCB (2000)
[2] SFB-RWQCB (1992)

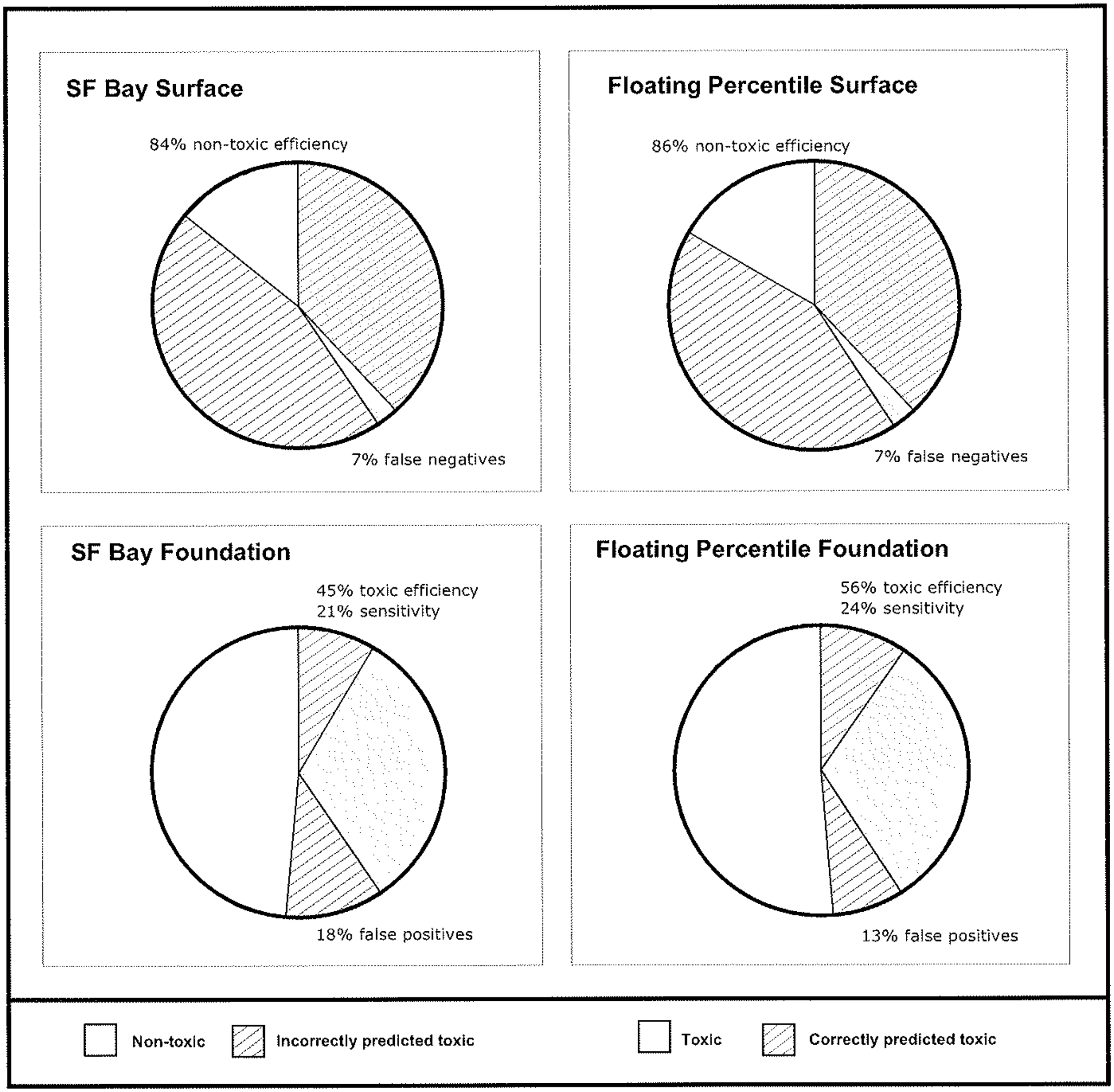
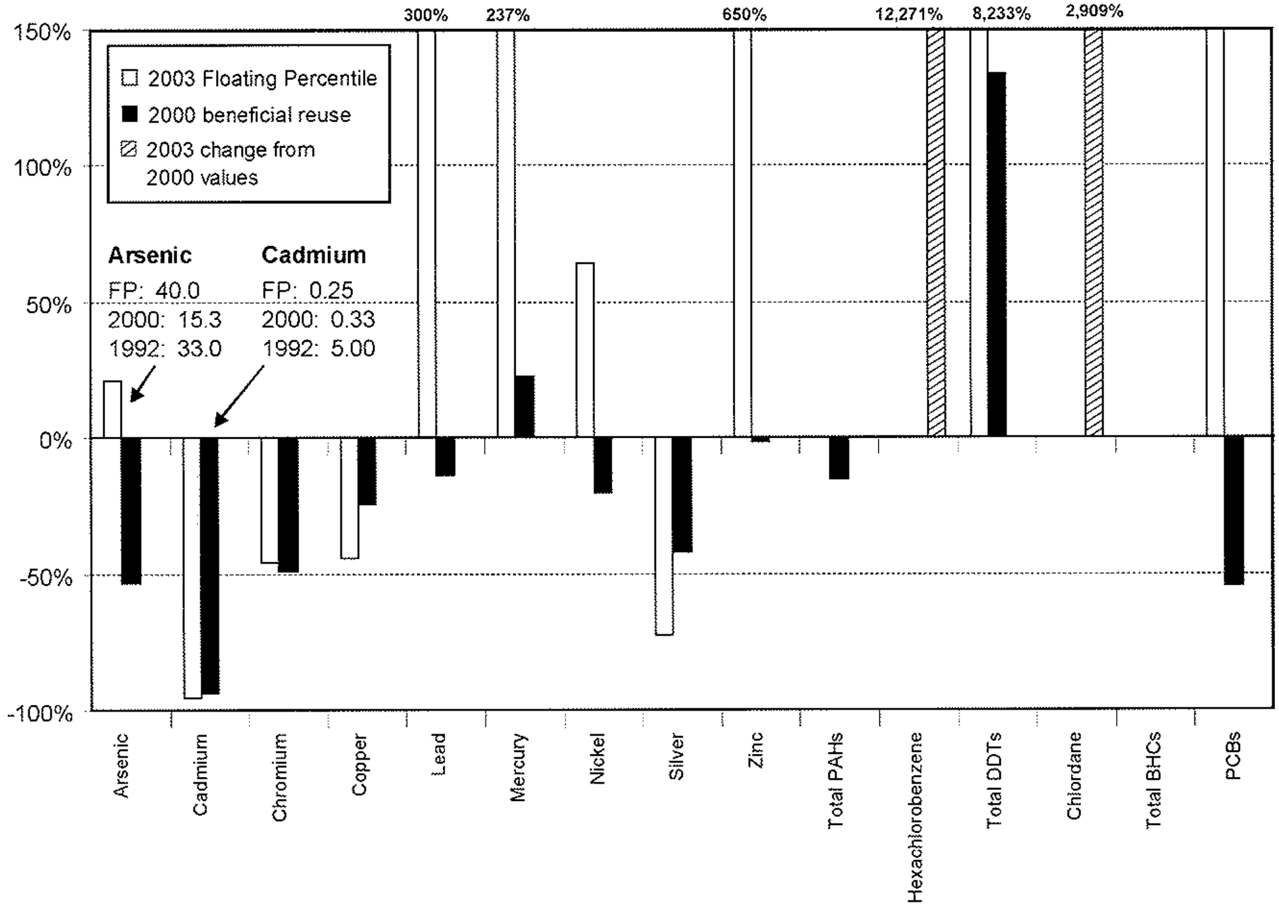
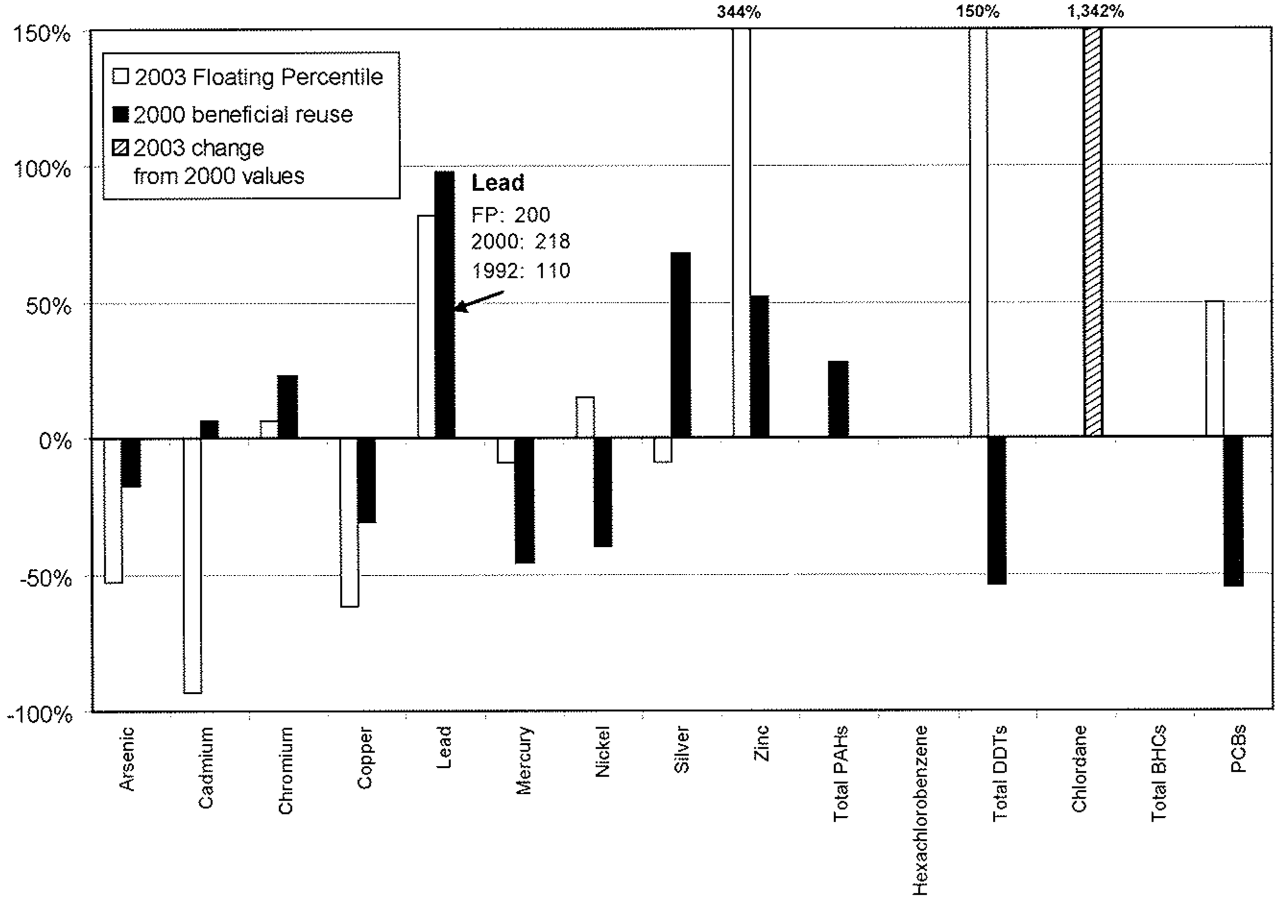


Figure 17. Performance evaluation results on complete data set for screening criteria for wetland surface and foundation material for the existing San Francisco Bay SQGs and newly developed site-specific Floating Percentile SQGs



Note: Values are expressed relative to 1992 screening criteria (SFB-RWQCB, 1992) when available.

Figure 18. Percent change in threshold concentrations for wetland surface screening criteria



Note: Values are expressed relative to 1992 screening criteria (SFB-RWQCB, 1992) when available.

Figure 19. Percent change in threshold concentrations for wetland foundation screening criteria

The FP process is a multivariate optimization routine that focuses first on the chemicals that are responsible for the most number of false positives, i.e., non-toxic samples that exceed the initial FP values. The final FP guidelines for chemicals that did not initially have a high number of false positives (e.g., arsenic, lead, mercury) are conditioned upon the optimized values for those chemicals that did (e.g., cadmium, chromium, total PAHs). For example, as you evaluate consecutively higher FP values for arsenic, you hold all other chemicals at their 'best-yet' thresholds. A higher FP value is selected only when it results in a lower false positive rate and a constant false negative rate, based on an evaluation of the complete chemical mixture relative to the full set of FP guidelines set to date. At concentrations above 40 ppm arsenic, all samples are toxic, whether from arsenic or the other chemicals in the mixture is unknown. At concentrations below 40 ppm arsenic, all non-toxic samples have chemical concentrations exceeding the other FP guidelines so they are not false negatives. Consequently, the FP guideline for arsenic goes right to a value of 40 ppm (the maximum concentration in non-toxic samples) and stays there, because below it, false negatives are not affected due to the other chemicals in the mixture exceeding their guidelines, while above it, false negatives would increase. When the targeted false negative rate changes from 5% to 75%, the same process is followed. Arsenic is again optimized after the other chemicals responsible for high numbers of false positives are optimized. The guideline again goes right to a value of 40 ppm because of the relationship of other chemicals in the mixture to the already optimized guidelines. In a sense, the toxicity threshold for arsenic in this dataset is well-defined, because it is constant across the range of false negative error rates.

The guideline values that change between surface and foundation are the chemicals that are associated with a large number of false positives. For example, the initial FP surface value for cadmium was associated with a large number of false positives, so the guideline for cadmium was raised as long as the false negative rate remained unchanged and the false positive rate decreased. If the guideline continues to increase for consecutively higher false negative rates, then the toxicity threshold is fuzzy: a higher guideline would increase the false negative rate, and a lower guideline would increase the false positive rate. Unlike the situation for arsenic, the non-toxic samples with concentrations below the guideline for cadmium are not predicted by other guidelines; therefore, the toxicity threshold is ambiguous.

Even though SSGs for many more chemicals than those listed in Table 12 have been established historically for both surface and foundation material (SFB-RWQCB, 1992, 2000), and no doubt the DMMO will (and should) continue to have permit applicants screen sediments for the same suite of contaminants that they normally require for sediment characterization (USACE/USEPA, 1999 a,b), the results from the exhaustive analyses presented in this document are clear: based on this data set, the only contaminants that can be reliably used for sediment screening guidelines are the fifteen chemicals listed in Table 12. It doesn't matter what the concentrations of selenium, for example, or benzo(a)pyrene are in a particular sediment as far as being able to predict an acute toxicity outcome in this particular data set (as evidenced by the summary distribution plots in Appendix C); the most reliable predictors for screening sediment samples in this data set are only the values presented in Table 12. How the values for these 15 chemicals can be used in a decision framework is presented in the final section of this report.

5.0 DISCUSSION

A noticeable difference between the suggested revisions to the SSG list (Table 13[2]) and the current and earlier regional guidelines (SFB-RWQCB, 1992, 2000) is that many chemicals (both metals and organic compounds) for which guideline threshold numbers for surface and foundation material have been established by the RWQCB are not included in the suggested list. Even though the number of contaminants in the suggested revised SSG list (Table 13) is much smaller than the standard suite of chemicals typically analyzed during permit testing, we are not advocating that the DMMO should change or reduce the number of analytes typically required for sediment characterization. The suggested revised guideline list is shorter because most of the chemicals measured in the historical regional data are not acute "toxicity drivers" or valid predictors of the bioeffects testing outcome; therefore, it is meaningless to have surface or foundation SSGs associated with any of these chemicals based on the historical data collected to date. The DMMO should keep collecting their standard suite of chemical data in the event that, as more data become available from a wider range of chemical concentrations, some clearer patterns may emerge for these "non predictors" (i.e., if toxic and non-toxic responses do not overlap completely for the chemical's measured range; a complete summary table of all the chemicals in the historical database and their concentration ranges is presented in Table 14). For any future application of the suggested SSGs in Table 13, special consideration would need to be given to surface material guidelines for the concentrations of mercury and persistent organic pollutants (POPs) such as hexacholorbenzene, DDTs, chlordane, and PCBs once a regional policy for dealing with bioaccumulative compounds of concern has been established.

Any regulatory agencies or stakeholders wanting to use the proposed revisions to the SSGs for surface and foundation material recommended in Table 13 should consider the policy decisions that were made at critical junctions during the course of work that affected the final calculations. A change to any of these decisions could dramatically affect the results of the performance evaluations as well as the final recommended guideline numbers; these policy decisions included:

- Choosing a 10% tolerance limit threshold for the *A. abdita* results (affects what chemical concentrations are associated with a toxic or non-toxic outcome)
- Choosing a 20% tolerance limit threshold for the *E. estuarius* results (also affects what chemical concentrations are associated with a toxic or non-toxic outcome)
- Choosing a false negative rate of 7% for surface material and a false positive rate of 13% for foundation material as acceptable

[2] Table 13 values do not account for bioaccumulation potential or chronic effects; regional guidelines for dealing with bioaccumulative compounds of concern have not yet been developed.

Table 13. Recommended sediment chemistry screening guidelines for beneficial reuse of dredged sediment in San Francisco Bay[1]

Chemical Name	Surface	Foundation
Metals (ppm, dry weight [DW])		
Arsenic[2]	40.0	40.0
Cadmium[3]	0.250	0.620
Chromium[3]	119	320
Copper[3]	50.0	150
Lead[3]	200	200
Mercury[2]	1.18	1.18
Nickel[2]	230	230
Silver[2]	0.280	2.00
Zinc[3]	1,200	1,200
Total PAH (molar sum)[4]	6.3	32
Chlorinated organic compounds (ppb, DW)		
Hexachlorobenzene[2]	60	60
Pesticides and PCBs (ppb, DW)		
Total DDTs[2]	250	250
Chlordane[4]	69.2	69.2
Total BHCs[4]	2.0	2.0
PCBs[2]	600	600

[1] Based on the Floating Percentile method for predicting acute amphipod toxicity

[2] Currently a Bioaccumulative Trigger (BT) in Puget Sound (PSDDA 2000).

[3] Proposed revision as a Bioaccumulative Contaminant of Concern in Puget Sound (BCOC; DMMO 2003).

[4] PAH BT/BCOC includes fluoranthene and benzo(a)pyrene; Chlordane BT includes only alpha-chlordane, BCOC includes total chlordanes; BHC BT includes only alpha-BHC, BCOC includes only gamma-BHC (lindane).

A change in what is considered acceptable for any of the above values could change the final guideline numbers in Table 13. What effect the tolerance limit threshold has on the reliability results and the final guideline numbers in Table 13 depends on the association between decreased survival and increased chemical concentrations. Using a higher survival threshold for defining toxic samples will result in non-toxic samples being reassigned as toxic samples. If the samples that change from non-toxic to toxic are associated with elevated concentrations in one or more chemicals, then using a higher toxicity threshold will improve the false positive and sensitivity rates (as was the case when the *E. estuarius* threshold was initially raised from the 10th to the 20th percentile), and at least some of the site-specific guidelines should decrease. With the database delivered with this report, the DMMO has the capability to see how changes in any of the above values affect both the final numbers and their screening performance in terms of sensitivity, reliability, toxic efficiency, and non-toxic efficiency.

Table 14: Summary overview of all chemicals in the historical database used for evaluations and analyses in this report

SSG Database Content Summary		All Chemistry Data					Detected Chemistry Data					Non-Detected Chemistry Data[1]					Summary		
Group/Chemical Name	Units	Count	Mean	Min[1]	Max	StdDev	Count	Mean	Min	Max	Std Dev	Count	Mean	Min	Max	Std Dev	Total	Total Detects	Total NDs
Metals																			
Aluminum	mg/kg	338	31105	1108 (B)	165000	23430	338	1108	165000	31105	23430						338	338	0
Antimony	mg/kg	125	3.94	0.623	19.5	2.69	58	0.62	19.50	2.72	3.60	67	5	5	5	0	125	58	67
Arsenic	mg/kg	524	12.88	2	1140	51.68	524	2	1140	12.88	51.68						524	524	0
Barium	mg/kg	67	121.97	25	240	36.88	65	52	240	124.95	33.17	2	25	25	25	0	67	65	2
Beryllium	mg/kg	67	0.62	0.25	1.7	0.33	41	0.53	1.7	0.85	0.20	26	0.25	0.25	0.25	0	67	41	26
Cadmium	mg/kg	607	0.408	0.033	27.9	1.45	579	0.033	27.90	0.41	1.49	28	0.05	0.85	0.29	0.32	607	579	28
Chromium	mg/kg	578	116.316	25.6	559	52.567	578	25.6	559	116.32	52.57						578	578	0
Cobalt	mg/kg	67	25	25	25	0						67	25	25	25	0	67	0	67
Copper	mg/kg	608	65.21	2.01	7,800	352.59	602	2.01	7800	65.80	354.30	6	2.29	25	6.07	9.27	608	602	6
Iron	mg/kg	318	35587	1799 (B)	336,000	29696	318	1799	336000	35587	29696						318	318	0
Lead	mg/kg	607	33.90	1.94	2100	102.26	607	1.94	2100	33.90	102.26						607	607	0
Manganese	mg/kg	277	629.13	49.6	2040	301.13	277	49.6	2040	629.13	301.13						277	277	0
Mercury	mg/kg	617	0.3034	0.0004 (B)	7.6800	0.4818	615	0.0004	7.6800	0.3043	0.4823	2	0.010	0.010	0.010	0	617	615	2
Methyl mercury	µg/kg	25	0.722	0.027	3.725	0.864	25	0.027	3.725	0.722	0.864						25	25	0
Molybdenum	mg/kg	67	25	25	25	0						67	25	25	25	0	67	0	67
Nickel	mg/kg	591	87.46	18.3	238	25.34	591	18.3	238	87.46	25.34						591	591	0
Selenium	mg/kg	606	0.419	0.024	35.727	1.518	565	0.030	35.727	0.434	1.569	41	0.024	0.750	0.201	0.244	606	565	41
Silicon	mg/kg	18	2.80	2.06	3.8	0.49	18	2.06	3.8	2.80	0.49						18	18	0
Silver	mg/kg	608	0.344	0.001	14.80	0.708	546	0.009	14.8	0.379	0.738	62	0.0006	0.7	0.034	0.089	608	546	62
Thallium	mg/kg	67	6.01	0.33	240	29.02	2	0.33	240	120.17	169.47	65	2.5	2.5	2.5	0	67	2	65
Tin	mg/kg	45	10.63	1.14	92.9	18.68	45	1.14	92.9	10.63	18.68						45	45	0
Vanadium	mg/kg	67	99.36	39	240	34.39	67	39	240	99.36	34.39						67	67	0
Zinc	mg/kg	608	148.19	9.92	6000	338.66	608	9.92	6000	148.19	338.66						608	608	0
PAHs																			
1-Methylnaphthalene	µg/kg	454	13.23	0.52	648	47.85	320	1.5	648	17.05	56.56	134	0.52	9.5	4.09	2.80	454	320	134
1-Methylphenanthrene	µg/kg	461	21.64	0.15	646	50.02	372	0.4	646	25.88	54.83	89	0.15	9.5	3.93	2.84	461	372	89
2,3,5-Trimethylnaphthalene	µg/kg	435	10.52	0.26	571.05	42.10	229	0.6	571.05	16.84	57.31	206	0.26	9.5	3.51	2.40	435	229	206
2,6-Dimethylnaphthalene	µg/kg	431	12.89	0.34	797.67	50.78	296	0.9	797.67	16.85	60.87	135	0.34	9.5	4.22	2.70	431	296	135
2-Methylnaphthalene	µg/kg	459	19.53	0.86	1010	63.93	349	1.8	1010	24.20	72.70	110	0.86	9.5	4.70	2.72	459	349	110
2-Methylphenanthrene	µg/kg	18	17.97	1.10	41.49	13.13	16	1.10	41.49	19.38	13.28	2	5.5	8	6.75	1.77	18	16	2

Table 14, continued

SSG Database Content Summary		All Chemistry Data					Detected Chemistry Data					Non-Detected Chemistry Data[1]					Summary		
Group/Chemical Name	Units	Count	Mean	Min[1]	Max	StdDev	Count	Mean	Min	Max	Std Dev	Count	Mean	Min	Max	Std Dev	Total	Total Detects	Total NDs
Acenaphthene	µg/kg	547	15.80	0.36	1350	69.74	407	0.8	1350	19.87	80.46	140	0.36	9.5	3.98	2.79	547	407	140
Acenaphthylene	µg/kg	547	22.09	0.21	1910	104.73	406	0.5	1910	28.46	120.94	141	0.21	9.5	3.73	2.77	547	406	141
Anthracene	µg/kg	565	51.14	0.39	3910	192.81	482	0.42	3910	59.26	207.70	83	0.39	9.5	3.95	2.49	565	482	83
Benz(a)anthracene	µg/kg	567	139.64	0.41	10900	512.16	542	0.75	10900	145.85	523.02	25	0.41	8	4.98	1.98	567	542	25
Benzo(a)pyrene	µg/kg	568	277.58	0.31	47300	2007.88	540	1	47300	291.74	2058.38	28	0.31	9	4.52	2.17	568	540	28
Benzo(b)fluoranthene	µg/kg	567	249.42	0.31	23200	1037.20	541	1	23200	261.19	1060.45	26	0.31	8	4.67	2.05	567	541	26
Benzo(e)pyrene	µg/kg	477	223.96	0.21	40600	1868.88	456	1	40600	234.04	1910.92	21	0.21	8	5.20	1.93	477	456	21
Benzo(g,h,i)perylene	µg/kg	568	244.89	0.31	32800	1397.11	545	1	32800	255.03	1425.44	23	0.31	7.5	4.62	1.84	568	545	23
Benzo(k)fluoranthene	µg/kg	566	96.41	0.28	5090	278.29	525	0.8	5090	103.66	287.71	41	0.28	8	3.58	2.38	566	525	41
Biphenyl	µg/kg	455	11.93	0.23	1360	65.74	346	0.7	1360	14.21	75.26	109	0.23	9.5	4.66	2.65	455	346	109
Chrysene	µg/kg	567	158.37	0.31	9990	507.18	540	1	9990	166.05	518.54	27	0.31	8	4.73	2.01	567	540	27
Coronene	µg/kg	39	416.14	5.46	9470	1511.68	39	5.46	9470	416.14	1511.68						39	39	0
Dibenz(a,h)anthracene	µg/kg	555	51.66	0.36	15500	658.78	449	0.71	15500	63.01	732.12	106	0.36	20	3.62	3.10	555	449	106
Fluoranthene	µg/kg	568	282.31	0.91	9300	590.55	545	0.91	9300	294.03	600.08	23	1.17	7.5	4.57	1.90	568	545	23
Fluorene	µg/kg	545	31.72	0.56	2330	155.29	412	0.8	2330	40.60	177.74	133	0.56	9.5	4.22	2.64	545	412	133
HPAH	µg/kg	551	2124.83	1	176020	8059.40	551	1	176020	2124.83	8059.40						551	551	0
Indeno(1,2,3-c,d)pyrene	µg/kg	567	178.78	0.31	7810	409.71	544	1	7810	186.14	416.69	23	0.31	7.5	4.77	1.89	567	544	23
LPAH	µg/kg	539	300.78	1	16121	892.48	539	1	16121	300.78	892.48						539	539	0
Methylanthracene	µg/kg	16	5.80	0.126	14.92	4.06	12	0.126	14.92	6.91	4.14	4	2.5	2.5	2.5	0	16	12	4
Naphthalene	µg/kg	551	27.98	0.6	806	58.88	445	1.1	806	33.59	64.25	106	0.6	20	4.45	2.87	551	445	106
PAHs	µg/kg	561	2426.53	1	177398.1	8408.11	561	1	177398.1	2426.53	8408.11						561	561	0
Perylene	µg/kg	459	92.05	1.2	9120	431.95	421	1.2	9120	100.00	450.22	38	2.5	24.4	4.05	3.61	459	421	38
Phenanthrene	µg/kg	567	130.42	0.62	7110	377.21	531	1	7110	138.97	388.32	36	0.62	8	4.20	2.01	567	531	36
Pyrene	µg/kg	567	388.21	1.03	22300	1128.04	545	1.03	22300	403.69	1147.93	22	2	7.5	4.73	1.79	567	545	22
Total PAHs (molar)	molar	557	12.09	0.01	918.84	43.45	557	0.01	918.84	12.09	43.45						557	557	0
Triphenylene	µg/kg	18	341.19	2.5	2760	639.04	16	17	2760	383.53	667.54	2	2.5	2.5	2.5	0	18	16	2
Pesticides																			
2,4-D	µg/kg	67	2647.02	1500	5460	981.87	9	3600	5460	4814.44	598.60	58	1500	3115	2310.69	457.33	67	9	58
2,4'-DDD	µg/kg	473	1.273	0.016	76.4	4.86	135	0.016	76.4	3.24	8.76	338	0.04	4.85	0.49	0.62	473	135	338
2,4'-DDE	µg/kg	474	0.540	0.017	14.1	0.91	41	0.017	14.1	0.88	2.28	433	0.05	2.75	0.51	0.64	474	41	433
2,4'-DDT	µg/kg	475	1.258	0.035	270	12.44	67	0.045	270	5.75	32.93	408	0.035	2.75	0.52	0.66	475	67	408
4,4'-DDD	µg/kg	569	4.740	0.045	310	20.85	407	0.045	310	6.37	24.47	162	0.05	2.75	0.63	0.57	569	407	162
4,4'-DDE	µg/kg	571	2.621	0.05	51.2	5.60	434	0.060	51.2	3.24	6.29	137	0.05	2.75	0.67	0.58	571	434	137

Table 14, continued

SSG Database Content Summary		All Chemistry Data					Detected Chemistry Data					Non-Detected Chemistry Data[1]					Summary		
Group/Chemical Name	Units	Count	Mean	Min[1]	Max	StdDev	Count	Mean	Min	Max	Std Dev	Count	Mean	Min	Max	Std Dev	Total	Total Detects	Total NDs
4,4'-DDT	µg/kg	572	2.148	0.023	163	8.73	259	0.023	163	4.02	12.73	313	0.05	2.75	0.60	0.63	572	259	313
Aldrin	µg/kg	531	0.421	0.025	3.06	0.62	18	0.100	3.06	0.53	0.65	513	0.025	2.75	0.42	0.62	531	18	513
alpha-BHC	µg/kg	550	1.011	0.002	292	12.49	63	0.002	292	5.32	36.85	487	0.035	2.75	0.45	0.64	550	63	487
alpha-Chlordane	µg/kg	488	1.199	0.008	51.2	3.67	123	0.008	51.2	3.06	6.90	365	0.05	5	0.57	0.74	488	123	365
alpha-Chlordene	µg/kg	47	1.474	0.16	11.5	2.50	24	0.160	11.5	2.65	3.10	23	0.25	0.25	0.25	0	47	24	23
beta-BHC	µg/kg	546	0.643	0.008	56.8	2.56	45	0.008	56.8	2.40	8.56	501	0.05	3	0.49	0.63	546	45	501
BHCs, total	µg/kg	118	4.967	0.005	456.62	42.26	118	0.005	456.62	4.97	42.26						118	118	0
Chlordane	µg/kg	257	5.359	0.035	171.37	17.87	189	0.035	171.37	6.81	20.67	68	0.2	1.695	1.33	0.23	257	189	68
cis-Nonachlor	µg/kg	387	0.509	0.006	18.7	1.48	87	0.006	18.7	1.28	2.86	300	0.035	2.75	0.29	0.51	387	87	300
Dacthal	µg/kg	60	0.307	0.1	11.1	1.42	4	0.216	11.1	3.21	5.27	56	0.1	0.1	0.10	0.00	60	4	56
DDTs	µg/kg	467	12.844	0.06	564.3	44.33	462	0.060	564.3	12.97	44.55	5	0.25	2.5	1.20	0.94	467	462	5
delta-BHC	µg/kg	535	0.662	0.05	99.4	4.36	11	0.100	99.4	10.63	29.74	524	0.05	2.75	0.45	0.61	535	11	524
Dieldrin	µg/kg	545	0.926	0.007	62.6	3.35	155	0.007	62.6	2.09	6.10	390	0.05	2.75	0.46	0.52	545	155	390
Endosulfan I	µg/kg	218	0.607	0.11	19.6	1.37	1	19.600	19.6	19.6		217	0.11	2.5	0.52	0.47	218	1	217
Endosulfan II	µg/kg	218	1.033	0.11	9.22	1.30	31	0.240	9.22	2.35	2.62	187	0.11	2.5	0.81	0.72	218	31	187
Endosulfan sulfate	µg/kg	218	1.993	0.11	163	11.19	18	0.570	163	15.24	37.24	200	0.11	10	0.80	0.86	218	18	200
Endrin	µg/kg	532	0.624	0.05	3.04	0.71	22	0.110	3.04	0.54	0.68	510	0.05	2.75	0.63	0.71	532	22	510
Endrin aldehyde	µg/kg	156	1.157	0.11	5	1.52						156	0.11	5	1.16	1.52	156	0	156
Endrin Ketone	µg/kg	67	1.072	0.6	2.5	0.41						67	0.6	2.5	1.07	0.41	67	0	67
Ethion	µg/kg	22	1.096	0.25	3.87	0.64	1	3.870	3.87	3.87		21	0.25	1	0.96	0.16	22	1	21
gamma-BHC	µg/kg	546	0.439	0.002	8.42	0.74	43	0.002	8.42	0.61	1.57	503	0.045	2.75	0.42	0.62	546	43	503
gamma-Chlordane	µg/kg	379	0.523	0.011	6.6	0.68	68	0.011	6.6	0.87	1.08	311	0.04	5	0.45	0.53	379	68	311
gamma-Chlordene	µg/kg	67	1.072	0.05	4.58	1.28	15	0.120	4.58	1.01	1.36	52	0.05	2.75	1.09	1.27	67	15	52
Heptachlor	µg/kg	530	0.999	0.025	12.5	1.90	31	0.031	7.26	0.97	1.49	499	0.025	12.5	1.00	1.92	530	31	499
Heptachlor Epoxide	µg/kg	550	1.060	0.004	17.8	2.04	26	0.004	17.8	1.08	3.43	524	0.02	12.5	1.06	1.95	550	26	524
Hexachlorobenzene	µg/kg	390	0.683	0.018	59.7	3.29	130	0.018	59.7	1.30	5.58	260	0.04	2.75	0.38	0.71	390	130	260
Methoxychlor	µg/kg	118	2.492	0.25	14.7	2.85	4	1.310	14.7	4.83	6.59	114	0.25	12.5	2.41	2.65	118	4	114
Mirex	µg/kg	438	1.426	0.05	103	5.27	52	0.180	103	2.87	14.18	386	0.05	12.5	1.23	2.13	438	52	386
Oxadiazon	µg/kg	60	4.704	0.25	114	14.36	4	0.850	114	29.25	56.50	56	0.25	3	2.95	0.37	60	4	56
Oxychlordane	µg/kg	387	0.343	0.006	2.75	0.63	16	0.006	2.57	0.41	0.67	371	0.02	2.75	0.34	0.63	387	16	371
p,p'-DDMS	µg/kg	36	1.5	1.5	1.5	0						36	1.5	1.5	1.5	0	36	0	36
p,p'-DDMU	µg/kg	75	1.85	0.10	30	4.24	29	0.09645	30	3.22	6.66	46	0.25	1	0.98	0.11	75	29	46
Silvex	µg/kg	67	397.95	225	830	148.82	8	650	830	755	66.33	59	225	460	349.53	68.58	67	8	59

Table 14, continued

SSG Database Content Summary		All Chemistry Data					Detected Chemistry Data					Non-Detected Chemistry Data[1]					Summary		
Group/Chemical Name	Units	Count	Mean	Min[1]	Max	StdDev	Count	Mean	Min	Max	Std Dev	Count	Mean	Min	Max	Std Dev	Total	Total Detects	Total NDs
Toxaphene	µg/kg	218	85.50	0.785	15700	1064.05	4	329.8	15700	4295	7604.22	214	0.785	37.5	6.82	7.41	218	4	214
trans-Chlordane	µg/kg	60	6.59	0.2	54.3	12.03	34	0.2	54.3	11.44	14.25	26	0.25	0.25	0.25	0	60	34	26
trans-Nonachlor	µg/kg	390	0.99	0.010	37.8	3.15	120	0.010	37.8	2.34	5.34	270	0.03	2.75	0.38	0.71	390	120	270
PCBs																			
PCBs	µg/kg	564	39.47	0.05	2183.32	140.24	434	0.1	2183.32	49.24	158.585	130	0.05	25	6.88	5.49	564	434	130
Phenols																			
Pentachlorophenol	µg/kg	67	398.97	225	830	148.95	8	650	830	755	66.33	59	225	475	350.69	69.72	67	8	59
Phthalates																			
Bis(2-ethylhexyl) phthalate	µg/kg	19	66.26	29	108	24.45	19	29	108	66.26	24.45						19	19	0
Butylbenzyl phthalate	µg/kg	19	5	5	5	0						19	5	5	5	0	19	0	19
Dibutyl phthalate	µg/kg	19	5	5	5	0						19	5	5	5	0	19	0	19
Diethyl phthalate	µg/kg	19	5	5	5	0						19	5	5	5	0	19	0	19
Dimethyl phthalate	µg/kg	19	5	5	5	0						19	5	5	5	0	19	0	19
Dioctyl phthalate	µg/kg	19	5	5	5	0						19	5	5	5	0	19	0	19
Phthalates	µg/kg	19	66.26	29	108	24.45	19	29	108	66.26	24.45						19	19	0
Other Semivols																			
Octamethyl pyrophosphoramide	µg/kg	47	7.33	0.019	67.6	13.88	47	0.019	67.6	7.33	13.88						47	47	0
p,p'-Dichlorobenzophenone	µg/kg	60	3.72	0.82	35.2	6.92	18	0.82	35.2	8.92	11.21	42	1.5	1.5	1.5	0	60	18	42
Dibenzothiophene	µg/kg	357	9.48	0.21	521	29.72	286	0.7	521	11.44	32.93	71	0.21	2.5	1.59	0.91	357	286	71
Phytane	µg/kg	68	31.40	0.52	144.07	28.59	67	1.9	144.07	31.86	28.55	1	0.52	0.52	0.52	0	68	67	1
Pristane	µg/kg	68	28.35	0.72	111.85	25.81	67	1.5	111.85	28.76	25.78	1	0.72	0.72	0.72	0	68	67	1
AVS/SEM																			
AVS	mg/kg	16	45.062	0.184	176	51.624	16	0.184	176	45.062	51.624						16	16	0
Cadmium SEM	mg/kg	16	0.016	0.0015	0.0451	0.013	16	0.0015	0.0451	0.016	0.013						16	16	0
Copper SEM	mg/kg	16	1.392	0.162	7.33	1.825	16	0.162	7.33	1.392	1.825						16	16	0
Lead SEM	mg/kg	16	0.797	0.0866	3.95	0.974	16	0.0866	3.95	0.797	0.974						16	16	0
Nickel SEM	mg/kg	16	0.281	0.137	0.512	0.095	16	0.137	0.512	0.281	0.095						16	16	0
SEM	mg/kg	16	10.118	1.53	31.3	8.666	16	1.53	31.3	10.118	8.666						16	16	0
SEM:AVS		16	1.377	0.0296	8.34	2.352	16	0.0296	8.34	1.377	2.352						16	16	0
Zinc SEM	mg/kg	16	7.631	0.725	28.3	7.294	16	0.725	28.3	7.631	7.294						16	16	0

Table 14, continued

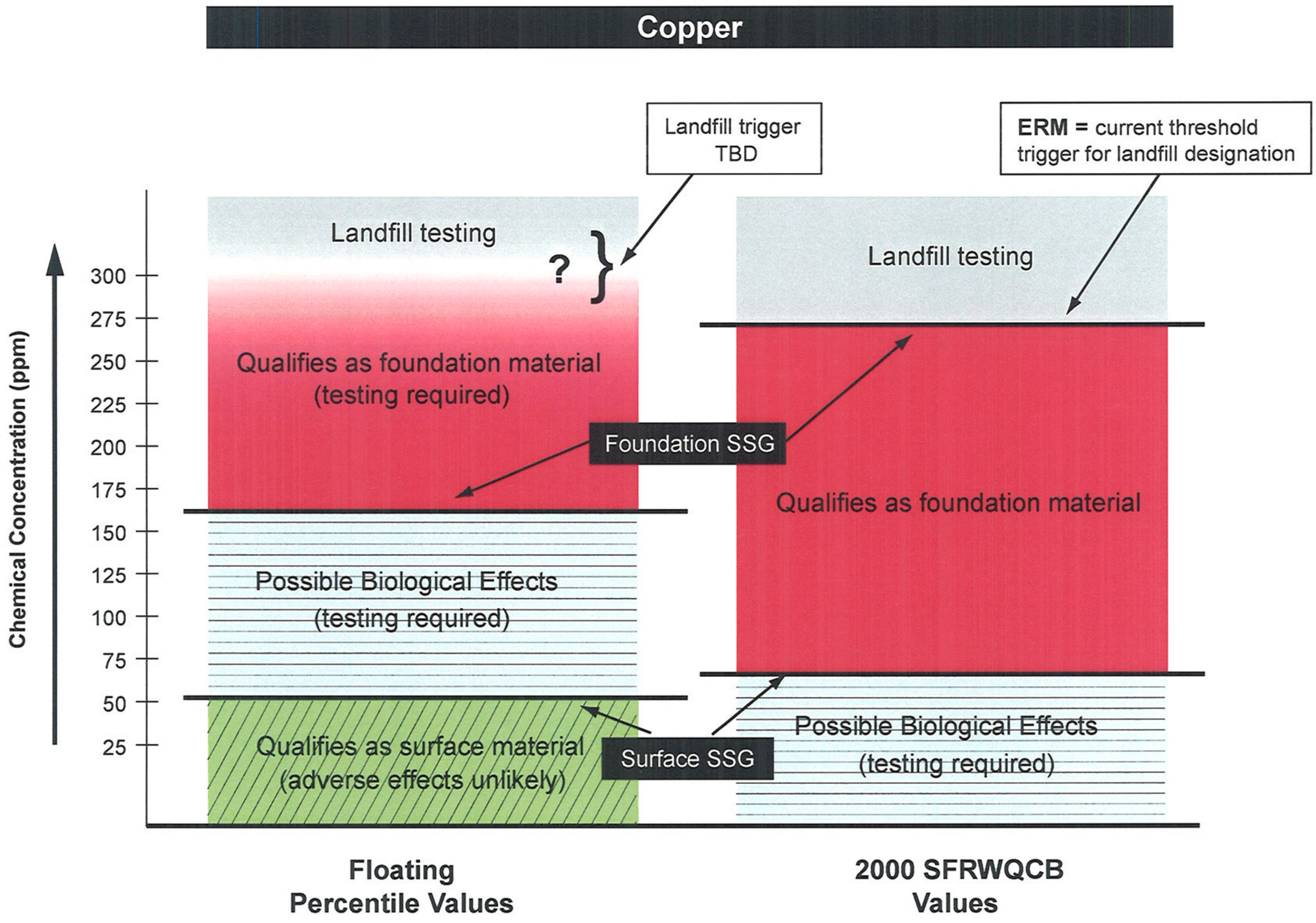
SSG Database Content Summary		All Chemistry Data					Detected Chemistry Data					Non-Detected Chemistry Data[1]					Summary		
Group/Chemical Name	Units	Count	Mean	Min[1]	Max	StdDev	Count	Mean	Min	Max	Std Dev	Count	Mean	Min	Max	Std Dev	Total	Total Detects	Total NDs
Butyltins																			
Butyltin	µg/kg	6	0.667	0.2	1.3	0.481	1	1.3	1.3	1.300		5	0.2	1.25	0.540	0.411	6	1	5
Dibutyltin	µg/kg	102	23.568	0.3	524	68.332	48	1.6	524	47.856	94.291	54	0.3	10.5	1.979	2.135	102	48	54
Monobutyltin	µg/kg	84	5.996	0.15	200	22.747	20	2.2	200	20.500	44.281	64	0.15	5	1.463	1.635	84	20	64
Tetrabutyltin	µg/kg	81	9.264	0.5	230	30.466	13	2.7	230	48.423	64.651	68	0.5	14	1.778	2.288	81	13	68
Tributyltin	µg/kg	202	47.250	0.09	5080	360.108	92	0.09	5080	101.728	530.015	110	0.25	6.5	1.686	2.216	202	92	110
Conventionals																			
Oil and Grease	mg/kg	72	138.89	50	720	108.60	44	110	720	195.45	105.13	28	50	50	50	0	72	44	28
pH		252	7.4	5.5	8.9	0.4	252.0	5.5	8.9	7.4	0.4						252	252	0
Salinity	ppt	26	19.05	4.9	33	9.11	26	4.9	33	19.05	9.11						26	26	0
Sulfides	mg/kg	158	287.46	0.05	1800	435.83	149	0.11	1800	304.82	442.91	9	0.05	0.5	0.1	0.15	158	149	9
Sulfides, dissolved	mg/l	224	0.25	0.0005	17	1.46	145	0.003	17	0.32	1.81	79	0.0005	0.495	0.13	0.13	224	145	79
Total Nitrogen	mg/kg	138	1365.22	200	3100	527.58	138	200	3100	1365.22	527.58						138	138	0
TOC	pct	688	1.27	0.05	7.47	0.81	681	0.08	7.47	1.28	0.80	7	0.05	0.05	0.05	0	688	681	7
TRPH	mg/kg	168	98.19	1.8	440	94.24	102	1.8	440	134.13	105.85	66	8	50	42.64	15.74	168	102	66
TVS	pct	154	4.03	0.45	7	1.58	154	0.45	7	4.03	1.58						154	154	0
Grain Size																			
Clay	pct	591	39.16	0.5	82	19.12	590	0.7	82	39.22	19.07	1	0.5	0.5	0.5		591	590	1
Gravel	pct	188	4.57	0.3	87	7.81	166	0.3	87	5.11	8.17	22	0.5	0.5	0.5	0	188	166	22
Sand	pct	515	31.87	1	99.2	28.22	515	1	99.2	31.87	28.22						515	515	0
Silt	pct	526	27.33	0.3	66	13.49	525	0.3	66	27.38	13.45	1	0.5	0.5	0.5		526	525	1

1 Values below detection reported as 1/2 the detection limit.

(B) Analyte found both in sample and associated blank.

A second important difference between the suggested foundation material SSGs recommended here and those currently in use (SFB-RWQCB, 2000) is that instead of these numbers being recommended as upper limits, they are instead minimum thresholds that would qualify a sediment to "cross over the line" as foundation material. Figure 20 illustrates the difference between the existing (SFB-RWQCB, 2000) and the newly-derived suggested SSG definitions. For the evaluation just completed for this project, sediments qualify for use as foundation material if any chemical concentration exceeds the suggested foundation SSGs; however, the permit applicant may decide to continue to invest resources in further testing protocols so that the sediment may eventually qualify for surface use (see tiered framework below). By design, foundation materials would not come in contact with any biological resources, so the cost of a wrong decision (using sediment that would be suitable for surface material as foundation material) may merely be a waste of potential clean sediment if it turns out that surface material is a limiting resource for a particular project.

The one thing we have not specified, as shown in Figure 20, is the upper limit for foundation material chemical concentrations that would prevent a sediment from being used as foundation material for any wetland restoration project and instead relegate it for landfill disposal. This is a regulatory, policy decision that would need to be established by the RWQCB; in the last iteration of their guidelines for wetland restoration (SFB-RWQCB, 2000), the RWQCB defaulted to ERM or PEL values as upper allowable limits for sediments to be used as foundation material. While upper limits do need to be established so that constructed or restored wetlands do not become Class I, II, or III waste management units, the current guidelines (ERM or PEL values) are extremely conservative given that any material placed as foundation material in wetlands would not come in contact with any biological receptors. There is a large difference between the landfill specific concentration values for material qualified for Class II (designated waste which could be released at concentrations in excess of applicable water quality objectives or which could cause degradation of state waters) waste management units and concentration values that exceed ERM or PEL thresholds; material could easily exceed ERM values (i.e., have chemical concentrations expected to elicit a toxic response) and not cause degradation to applicable water quality objectives. In other words, it is possible for sediments with concentrations above ERM values to have the same characteristics of sediments that would fail biological testing in the area shown in Figure 20 labeled, "Possible Biological Effects – Testing Required". There is a great deal of latitude that could be exercised to increase the volume capacity for foundation material while not sacrificing or compromising environmental quality. Given the cost differential between landfill disposal and wetland construction, the increased potential for environmental problems associated with taking marine sediments in upland settings, and the LTMS goals for 40% of all dredged material to go toward beneficial re-use, it appears that it would be a very worthwhile effort for the appropriate regulatory agencies to re-examine the current guidelines for upper threshold limits of wetland foundation material and possibly establish new policy guidelines for foundation sediment upper concentration limits.



Copper
Landfill trigger TBD
ERM = current threshold trigger for landfill designation
Chemical Concentration (ppm)
300
275
250
225
200
175
150
125
100
75
50
25
Landfill testing
?
Qualifies as foundation material (testing required)
Foundation SSG
Possible Biological Effects (testing required)
Qualifies as surface material (adverse effects unlikely)
Surface SSG
Landfill testing
Qualifies as foundation material
Possible Biological Effects (testing required)
Floating Percentile Values
2000 SFRWQCB Values

Figure 20. Differences in how SSG thresholds are defined in this report vs. the SFRWQCB 2000 report

A third importance difference between the suggested surface material guidelines in Table 13 and the current ones is that sediments that qualify as surface material can be used without having to undergo toxicity testing. Because the predictive validity of these SSGs are known, permit applicants would not have to allocate resources to unnecessary test procedures; the resource management implications of implementing these suggested guidelines are discussed in the next section.

6.0 CONCLUSIONS AND RECOMMENDATIONS

Now that a set of suggested screening guidelines is available based on regional data with known reliability and toxic/non-toxic efficiency (Table 13), it is appropriate to revise the tiered testing framework (for the two earlier iterations, see SFB-RWQCB, 1992, 2000) so that the suggested SSGs can be used in a proper context for resource management. Figure 21 shows a proposed tiered testing framework that matches the level of required testing to the management goals and environmental protection associated with reuse of dredged material for either wetland projects (surface or foundation material) or landfill allocation. As with any general framework, site-specific factors or restoration goals should be taken into consideration when evaluating material to be used. There are five major differences in the framework shown in Figure 21 as compared with earlier versions if these suggested guidelines are adopted:

1. Surface material guidelines act as true guidelines; if none are exceeded, the material is suitable for wetland cover with no required additional testing (as long as bioaccumulation triggers are not exceeded).

2. Foundation material upper limits ("landfill trigger" in Figure 20) need to be established by the appropriate regulatory agencies.

3. The potential for bioaccumulation is recognized an integral early decision tier for surface material considerations in this framework; however regional bioaccumulation triggers have not been established as yet and should be a high priority for the DMMO agencies.

4. Permit applicants are given a choice (the "dredger's option") to subject the material to additional bioeffects testing instead of accepting the uncertainty associated with bioaccumulation trigger (BT) values or material with concentrations above surface SSGs.

5. This framework recognizes that bulk sediment chemistry values alone do not necessarily imply contaminant bioavailability; just as the DMMO has recognized in the past that concentrations well in excess of published SSGs for nickel and chromium are not associated with adverse effects, this framework allows the permit applicant to pursue additional bioeffects testing to verify the suitability of sediment for use as wetland surface material.

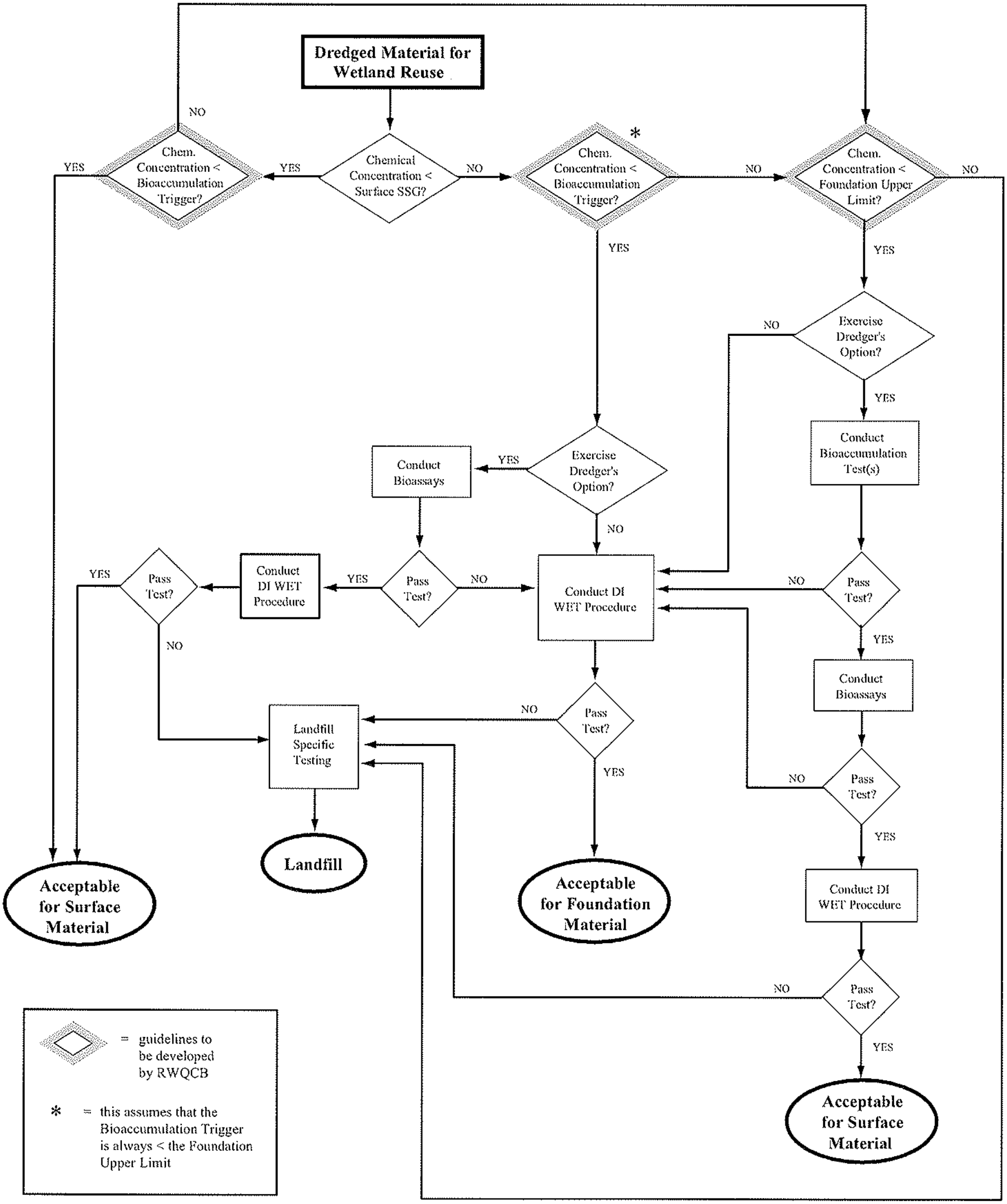

Dredged Material for Wetland Reuse
Chemical Concentration < Surface SSG?
YES
NO
Chem. Concentration < Bioaccumulation Trigger?
NO
YES
*
Chem. Concentration < Bioaccumulation Trigger?
NO
YES
Chem. Concentration < Foundation Upper Limit?
NO
YES
Exercise Dredger's Option?
NO
YES
Conduct Bioaccumulation Test(s)
Exercise Dredger's Option?
YES
NO
Conduct Bioassays
Pass Test?
YES
NO
Conduct DI WET Procedure
Pass Test?
YES
NO
Conduct DI WET Procedure
Pass Test?
NO
YES
Conduct Bioassays
Pass Test?
NO
YES
Pass Test?
NO
YES
Landfill Specific Testing
Landfill
Acceptable for Surface Material
Acceptable for Foundation Material
Conduct DI WET Procedure
Pass Test?
NO
YES
Acceptable for Surface Material
= guidelines to be developed by RWQCB
* = this assumes that the Bioaccumulation Trigger is always < the Foundation Upper Limit

Figure 21. Proposed tiered testing framework for dredged material reuse or disposal

The most radical departure from earlier versions of this tiered framework is the incorporation of BT values as a decision point. Because the final suggested SSGs are based on acute effects testing results, some of the recalculated SSG values for DDTs (250 ppb) and PCBs (600 ppb) are clearly above generally accepted safe thresholds. These particular persistent organic pollutants are included on the United Nations Environment Program list of POPs selected for global action (Rodan et al., 1999); the concern associated with POPs is chronic, not acute effects. Given the occurrence of POPs in San Francisco Bay sediments, any sediment management framework adopted for resource management would need to take their effects into consideration.

Because regional BT values for the San Francisco Bay area still need to be developed by the DMMO, we would urge that resources be allocated to tackle this task as soon as possible; the management framework in Figure 21 cannot be implemented until policy decisions are made by the regional regulatory agencies concerning both ceiling thresholds for foundation material and appropriate bioaccumulation triggers for surface material. As with previous wetland restoration guidance documents (SFB-RWQCB, 1992, 2000), the final evaluation of foundation material may be performed using a modified Waste Extraction Test using de-ionized water (designated as DI WET in Figure 21), as defined in Title 23 of the California Code of Regulations; any water discharged during material placement would also need to be evaluated by both chemical analyses and biological tests for elutriate toxicity (USACE/USEPA 1999a).

While all the metals listed in Table 13 have the potential for bioaccumulation, mercury (Hg), is one of the most prominent as well as the most problematic in terms of bioaccumulation. Mercury contamination in the Bay area is a serious problem resulting from historic mining sites in the Sacramento and San Joaquin river watersheds. The California Bay-Delta Authority (CALFED) recently developed a Mercury Science Strategy (Strategy) to provide an integrated framework for evaluating mercury contamination in the Bay-Delta System, and to link these investigations to restoration projects (Wiener et al. 2003). The goals of the Strategy are (1) to assist and recover at-risk native species, (2) to rehabilitate the Bay-Delta to support native aquatic and terrestrial biotic communities, (3) to maintain or enhance selected species for harvest, (4) to protect and restore functional habitat for both ecological and public values, (5) to prevent the establishment of additional non-native species, and (6) to improve or maintain water and sediment quality.

Although the Strategy does not cite any sediment Hg contaminant concentration goals, a sediment TMDL was recently proposed for the State of California (Johnson and Looker 2003). The sediment Hg target was considered preferable to a water quality target, because sediment concentrations are directly related to Hg in the Bay and are less subject to short-term fluctuations. The target was generated from data collected through the Regional Monitoring Program (RMP) from 1993 to 2000. The report states that to meet the proposed fish tissue and bird egg targets, a 40 to 50% reduction is needed in the amount of Hg in San Francisco Bay sediment. A final median sediment Hg concentration of 0.2 ppm has been proposed as the sediment mercury target (Johnson and Looker 2003).

While the new suggested mercury SSG for both surface and foundation at 1.18 ppm is about six times the proposed target concentration of Johnson and Looker (2003), the real concern about bioaccumulation with mercury occurs once it becomes methylated; however, the processes controlling mercury cycling and methylation are poorly understood. Fate and transport of mercury in a wetland environment is dependent on the chemical form, or valence state. The divalent form of Hg (Hg^{+2}) can combine with both inorganic and organic compounds, and can therefore be converted to methyl mercury (MeHg), the most bioavailable, and therefore toxic form. Elemental mercury (Hg^{0}) is rare in the environment, and not available for methylation. Inorganic, or monovalent Hg (Hg^{+1}) is the most common form released to the environment and bound to particulates. Inorganic Hg combines with inorganic compounds only, it cannot be methylated, and is therefore much less bioavailable; a brief summary of the factors affecting mercury bioaccumulation is presented in Appendix E.

The other components shown in the tiered testing framework (Figure 21) are familiar elements from other regional dredged material framework guidance documents and do not require further explanation. While the tiered framework presented in Figure 21 presents a logical structure for incorporating the suggested revisions to the wetland SSGs in a resource management framework, project proponents must always recognize that any proposed project or sampling/testing design needs prior approval from the DMMO. This would include acceptable sampling program designs, sampling and analysis plans, and reporting requirements as outlined in Public Notice 99-4 and 01-01 (USACE/USEPA 1999, DMMO 2001).

The results from this project allow regulators and permit applicants to evaluate the suitability of dredged material for various disposal/reuse alternatives based on SSGs that were calculated on regional data and that have a known reliability performance. With the associated final database deliverable that accompanies this report, DMMO representatives have the option of updating these suggested guidelines periodically as more data become available or if regulatory consensus changes over time regarding policy decisions about acceptable false positive or false negative rates for surface and foundation material guidelines.

Both the distribution plots and ROC curves (Figures C-1 and C-6, Appendix C) dramatically illustrate why there were clear limits to the sensitivity, specificity, and predictive reliability of the suggested guidelines developed from the existing data: both the toxic and non-toxic responses overlap one another for most of the range of chemical concentrations measured. If the suggested guidelines are to be revised or updated in the future, the question of when it would be worthwhile to do so will not be dependent on merely increasing the number of data points in the database, but increasing the range of chemical concentrations represented in the database (Table 14) so that hopefully there would be some clear separation between toxic and non-toxic response. That will not only improve the chances of achieving lower false positive or false negative rates, as well as higher reliability, but also allow surface and foundation material guidelines to be calculated for contaminants not presently included on the revised list in Table 13.

Data from the San Francisco Bay area as well as other areas in California are being collected for a statewide sediment quality objective development program (Bay et al. 2003), and will include dredging programs such as the Port of Oakland 38 and 42 foot projects, and the Richmond Harbor Deepening project. Addition of these samples including relatively contaminated sediments with paired bioassay results could help to reduce the uncertainty in the analyses presented here, but would most likely not eliminate the uncertainty. The database will also include bioaccumulation and benthic infaunal data, however, providing alternate endpoints that could aid in a weight-of-evidence approach in developing sediment screening guidelines.

Our analysis for the performance evaluations of regional and national guidelines on the existing data has been exhaustive, and we feel quite confident that we have calculated the optimal SSGs within the limits of the regional data that are presently available. The suggested surface and foundation SSGs listed in Table 13 provide the DMMO with the best available values based on the policy decisions made during the course of our investigations. The tiered framework provided in Figure 21 integrates these revised SSGs in a logical approach for their application to future permitting decisions. As with any guidelines, these are based on the best available information at the present time and therefore should be considered interim; they are subject to future improvement as our state of knowledge increases, data from additional bioassay species and/or sublethal endpoints are found, or more information becomes available to justify recalculation of the reference tolerance limits or the SSGs.

This report has demonstrated both the uncertainty as well as the predictive validity associated with the single line of evidence (amphipod acute toxicity test results) most commonly applied to sediment assessment projects. However, it is important to keep in mind that, at present, there exists no one "perfect" toxicity test or sediment assessment method. The amphipod test is one of the most widely applied for marine sediment assessment programs, and while the "perfect test" will probably never be developed, one should not abandon existing predictors because they are fallible (Arkes et al., 1986). We feel that future efforts toward improved regulatory decision-making in the sediment arena should not focus so much on developing a better toxicity test, but instead focus on utilizing multiple lines of evidence with known predictive validity that best integrate information about the physical, chemical, and biological attributes of the material under consideration; the tiered framework in Figure 21 provides guidance on where resources should be allocated toward that objective. Lines of evidence that need to be evaluated for wetland restoration projects include leachate characteristics of the sediments under both aerobic and anaerobic conditions as well as bioaccumulation potential for material in surface layers that comes in contact with biological receptors.

While recent studies have provided some insights into the mechanisms of mercury methylation in San Francisco Bay sediments (Marvin-DiPasquale & Agee, 2003), there clearly remains a great deal of research to be done toward understanding the processes and therefore being able to predict conditions under which mercury methylation will occur. Even though researchers will no doubt continue to develop new chronic or acute bioassay tests, both applied ecologists and resource managers will be forced to deal with the wide range of variability that exists in nature. Regardless of whether or not these suggested SSGs are adopted by regional regulators, it is highly unlikely that any one set of numbers would ever provide all the guidance needed for sediment beneficial re-use projects; there will always be exceptions to any rules developed. Any sediment guidelines used for regulatory purposes are best employed in the context of a resource management framework similar to the one in Figure 21 where additional lines of evidence are taken into consideration.

Given the amount of variability in all environmental data sets and the amount of uncertainty inherent in our knowledge of ecosystem function (Germano, 1999), resolving any environmental issue always entails more than finding a technical or analytical solution; environmental decisions reflect politics, social and cultural values, and expectations, as much as scientific facts (Bardwell, 1991). We rarely encounter risk-free decisions, so regulatory decisions ultimately depend on what priorities and trade-offs stakeholders choose to accept. Resource managers in charge of sediment projects will always face the dilemmas of dealing with uncertainty and natural variability, and the best option for managing uncertainty is to use a weight of evidence approach in the decision-making process. While employing multiple lines of evidence will decrease the reliance on one set of imperfect guidelines, regulators and resource managers also need to avoid the common trap of employing too many lines of evidence with unknown predictive validity. While it is quite common to feel that having more information will lead to reduced uncertainty and a better (or more informed) decision, studies have shown that actually the opposite is true (Germano, 1999). Accurate diagnosis is best achieved with a limited set of valid predictors (ca. 3 or 4), and improving judgmental accuracy is usually more an exercise in exclusion than one of inclusion (Faust 1989). The same investigative rigor that has been applied to this one line of evidence to assess its predictive validity would also need to be applied to any other lines of evidence under consideration in a sediment regulatory framework to determine whether their inclusion would actually help or hinder the final decision-making process.

To ensure progress toward implementing the management framework in Figure 21 (or a similar revision based on updated information), our final recommendations are as follows:

- The DMMO should continue to maintain and augment the assembled database to overcome the present limitations (too much overlap between toxic and non-toxic responses because of limited results from highly contaminated material); a re-evaluation of these suggested SSGs may be warranted in the future once more data are available with greater concentration ranges than those that presently exist in the database.

- Regulatory efforts should be devoted in the near term to developing regional bioaccumulation triggers; any BT guidelines established should be validated against historical bioaccumulation testing data similar to what has been done in this report with toxicity testing data to see if they are achieving the intended or desired reliability (this would also require establishment and maintenance of a database for regional bioaccumulation test results).

- Revised ceiling limits for foundation material should be re-examined if the suggested SSGs are adopted; the existing foundation material ceiling limits are overly conservative, had poor reliability results (Section 3.2.2), and do not take advantage of the isolation that foundation material would have from any biological receptors.

- Both state and federal regulators should recognize the serious limitations with the *E. estuarius* data pointed out in our Task 4.1 memorandum (Appendix B) and consider the implications for continued use of this species on future SSG development. If *E. estuarius* continues to be used for testing, special attention should be paid to controlling for confounding factors (Appendix B; Word et al., In Press).

- Wetland restoration projects should include a long-term monitoring component to provide feedback verification that the guidelines being used by the regulatory agencies are indeed achieving the level of environmental protection desired.

7.0 REFERENCES

Arkes, H.E., R.M. Dawes, and C. Christensen. 1986. Factors influencing the use of a decision rule in a probabilistic task. *Behavioral Human Decision Processes* 37: 93-110.

ASTM. 1998. Standard guide for conducting 10-day static sediment toxicity tests with marine and estuarine amphipods. E1367-92. Volume 11.05: 732-757. Annual Book of Standards: American Society of Testing and Materials, Conshohocken, PA.

Bagui, S.C., D.K. Bhaumik, and M. Parnes. 1996. One-sided tolerance limits for unbalanced M-Way Random-Effects ANOVA Models. *Journal of Applied Statistical Science*, 2/3:135-148.

Bardwell, L.V. 1991. Problem-framing: a perspective on environmental problem-solving. *Environmental Management,* 15:603-612.

Bay, S., D. Vidal, and C. Beegan. 2003. Revised workplan for the development of sediment quality objectives for enclosed bays and estuaries of California. State Water Resources Control Board, Resolution No. 2003-2004. Available at: http://www.swrcb.ca.gov/bptcp/docs/finalworkplan052103.pdf.

Chapman, P.M. 2000. Why are we still emphasizing chemical screening-level numbers? *Marine Pollution Bulletin,* 40:465-466.

Clarke, J, C. Lutz, and V. McFarland. 1988. Influence of environmental variables on bioaccumulation of mercury. US Army Waterways Experiment Station Technical Environmental Effects of Dredging Technical Note EEDP-01-14. WES, Vicksburg, MS.

Connell, D. and R. Markwell. 1992. Mechanism and prediction of nonspecific toxicity to fish using bioconcentration characteristics. *Ecotoxicology and Environmental Safety,* 24:247-265.

CSWRCB. 1998. Sediment quality and biological effects in San Francisco Bay. Bay Protection and Toxics Cleanup Program. California State Water Resources Control Board. Final Technical Report, August 1998.

Dewitt, T.H., R.C. Swartz, and J.O. Lamberson. 1989. Measuring the toxicity of estuarine sediment. *Environmental Toxicology and Chemistry*, 8:1035-1048.

DiToro, D.M., J.A. McGrath, and D.J. Hansen. 2000. Technical basis for narcotic chemicals and polycyclic aromatic hydrocarbon criteria. I. Water and Tissue. *Environmental Toxicology and Chemistry*, 19: 1951-1970.

DMMP. 2003. Revisions to the Bioaccumulative Contaminants of Concern (BCOC) List. DMMP Issue Paper, prepared by Erika Hoffman, USEPA, for the DMMP agencies. Presented at SMARM (Sediment Management Annual Review Meeting), May 7, 2003. http://www.nws.usace.army.mil/publicmenu/DOCUMENTS/BCOC.pdf

DMMO. 2001. Guidelines for Implementing the Inland Testing Manual in the San Francisco Bay Region. September 21, 2001. http://www.spn.usace.army.mil/conops/sfitm092101.pdf

Efron, B. and R.J. Tibshirani. 1993. An introduction to the bootstrap. Monographs on statistics and applied probability 57. Chapman and Hall, New York. 436 pp.

Fairey, R., C. Bretz, S. Lamerdin, J. Hunt, B. Anderson, S. Tudor, C. Wilson, F. LaCaro, M. Stephenson, M. Puckett, and E. Long. 1996. Chemistry, toxicity, and benthic community conditions in sediments of the San Diego Bay region. California State Water Resources Control Board, Sacramento, CA.

Fairey, R., E.R. Long, C.A. Roberts, B.S. Anderson, B.M. Phillips, J.W. Hunt, H.R. Puckett, and C.J. Wilson. 2001. An evaluation of methods for calculating mean sediment quality guideline quotients as indicators of contamination and acute toxicity to Amphipods by chemical mixtures. *Environmental Toxicology and Chemistry*, 20: 2276-2286.

Faust, D. 1989. Data integration in legal evaluations: Can clinicians deliver on their premises? Behav. Sci. Law 7: 469-483.

Field, L. J., D. MacDonald, S. B. Norton, C. G. Ingersoll, C. Severn, D. Smorong, and R. Lindskoog. 2002. Predicting amphipod toxicity from sediment chemistry using logistic regression models. *Environmental Toxicology and Chemistry*, 21(9):1993-2005.

Gandesbery, T. F. Hetzel, R. Smith, and L. Riege. 1998. Ambient concentrations of toxic chemicals in San Francisco Bay Sediments: summary. San Francisco Regional Water Quality Control Board Staff Report, May 1998. http://www.sfei.org/rmp/1997/c0405.htm#a06

Germano, J.D. 1999. Ecology, statistics, and the art of misdiagnosis: The need for a paradigm shift. *Environmental Reviews,* 7:167- 90.

Hunt, J. W., B. S. Anderson, B. M. Phillips, J. Newman, R. Tjeerdema, M. Stephenson, M. Puckett, R. Fairey, R. W. Smith, and K. Taberski. 1998. Evaluation and use of sediment reference sites and toxicity tests in San Francisco Bay. Prepared for California State Water Resources Control Board. Can be obtained from: http://www.swrcb.ca.gov/general/publications/index.html#Ss

Johnson, B. and R. Looker. 2003. Mercury in San Francisco Bay, Total Maximum Daily Load (TMDL) Project Report. California Regional Water Quality Control Board, San Francisco Bay Region.

http://www.swrcb.ca.gov/rwqcb2/TMDL/SFBayMercury/SFBayMercuryTMDLProjectReport.pdf

Landrum, P.F., G.R. Lotufo, D.C. Gossiaux, M.L. Gedeon, and J.H. Lee. 2003. Bioaccumulation and critical body residue of PAHs in the amphipod, *Diporeia* spp.: additional evidence to support toxicity additivity for PAH mixtures. *Chemosphere*. 51:481-489.

Long, E.R. and D.D. MacDonald. 1992. National Status and Trends Program Approach. In: Sediment Classification Methods Compendium. EPA 823-R-92-006. EPA Office of Water (WH-556), Washington, DC

Long, E.R., D.D. MacDonald, S.L. Smith, and F.D. Calder. 1995. Incidence of adverse biological effects within ranges of chemical concentrations in marine and estuarine sediments. *Environmental Management,* 19:81-97.

McBride, G. B. 1999. Equivalence tests can enhance environmental science and management. *Austral. & New Zealand J. Statist.*, 41:19-29.

MacDonald, D.D., R.S. Carr, F.D. Calder, and E.R. Long. 1995. Development and evaluation of sediment quality guidelines for Florida coastal waters. *Ecotoxicology.*

Marvin-DiPasquale, M., and J.L. Agee. 2003. Microbial mercury cycling in sediments of the San Francisco Bay-delta. *Estuaries* 26: 1517-1528.

Nichols, F. H., J.E. Cloern, S.N. Luoma, and D. H. Peterson. 1986. The modification of an estuary. *Science,* 231:525-648.

O'Connor, T.P. and J.F. Paul. 2000. Misfit between sediment toxicity and chemistry. *Marine Pollution Bulletin*, 40:59-64.

PSDDA. 2000. Dredged material evaluation and disposal procedures. A user's manual for the Puget Sound Dredged Disposal Analysis (PSDDA) Program. USACE, USEPA, WADNR, WA Ecology, February 2000. http://www.nws.usace.army.mil/publicmenu/Attachments/UMPDF.pdf

Rodan, B.D., D.W. Pennington, N. Eckley, and R.S. Boethling. 1999. Screening for persistent organic pollutants: Techniques to provide a scientific basis for POPs criteria in international negotiations. *Environ. Sci. Technol.*, 33:3482-3488.

SAIC/Avocet. 2002. Draft Report. Development of freshwater sediment quality values for use in Washington State, Phase I Task 6: Final Report. September 2002. Publication Number 02-09-050. Prepared for Washington Department of Ecology, Sediment Management Unit. Prepared by Science Applications International Corporation and Avocet Consulting.

SFB-RWQCB. 1992. Sediment screening criteria and testing requirements for wetland creation and upland beneficial reuse. Interim Final. Public Notice No. 92-145.

SFB-RWQCB. 2000. Beneficial reuse of dredged materials: sediment screening and testing guidelines. Draft Staff Report. May, 2000.

Shine, J.P., Trapp, C.J, and B.Z. Coull. 2003a. Use of receiver operating characteristic curves to select and evaluate sediment quality guidelines. http://www.hsph.harvard.edu/water/ROCpresentation.pdf.

Shine, J.P., Trapp, C.J., and B.Z. Coull. 2003b. Use of receiver operating characteristic curves to evaluate sediment quality guidelines for metals. *Environmental Toxicology and Chemistry,* 22 (7):1642-1648.

Smith, R. W. 2002. The use of random-model tolerance intervals in environmental monitoring and regulation. *Journal of Agricultural, Biological and Environmental Statistics,* 7(1): 74-94.

Smith, R. W. and L. Riege. 1999. San Francisco Bay sediment criteria project ambient analysis report. Prepared for California Regional Water Quality Control Board, San Francisco Bay Region by EcoAnalysis, Inc., Ojai, CA.

Tetra Tech. 1986. Development of sediment quality values for Puget Sound. Task 6. Final Report. Prepared for Resource Planning Associates for U.S. Army Corps of Engineers. Tetra Tech, Inc., Bellevue, WA.

USACE. 1998. Long term management strategy (LTMS) for the placement of dredged material in the San Francisco Bay Region. Final. EIS/EIR. Published jointly by USACE, USEPA, BCDC, SFB-RWQCB, and SWRCB.

USACE/USEPA. 1999. Sampling and analysis plan (quality assurance project plan) guidance for dredging for dredging projects within the San Francisco District. July 1999. Public Notice No. 99-4.

USACE/USEPA/WDNR/WDOE. 2000. Dredged material evaluation and disposal procedures. a users manual for the Puget Sound Dredged Disposal Analysis (PSDDA) Program. 88 pages. http://www.nws.usace.army.mil/publicmenu/Attachments/UMPDF.pdf

USEPA. 1992a. A Supplemental Guidance to RAGS: Calculating the Concentration Term. http://www.deq.state.ms.us/newweb/opchome.nsf/pages/HWDivisionFiles/$file/uclmean.pdf, Publication 9285.7-081. Office of Solid Waste and Emergency Response, U.S. Environmental Protection Agency, Washington, D.C.

USEPA. 1992b. Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities: Addendum to Interim Final Guidance. Office of Solid Waste, U.S. Environmental Protection Agency, Washington, D.C. Currently available as part of: Statistical Training Course for Groundwater Monitoring Data Analysis, EPA/530-R-93-003.

USEPA. 2001. Region 10: Inventory of persistent, bioaccumulative, and toxic (PBT) chemicals reported to the toxic release inventory 1991 – 1998. USEPA Region 10, May 2001. http://yosemite.epa.gov/r10/owcm.nsf/0d511e619f047e0d88256500005bec99/6ad9c10eb8a06bc288256506007def78/$FILE/pbt98_2.PDF

Veith, G.D., D.J. Call, and L.T. Brooke. 1983. Structure-Toxicity Relationships for the Fathead Minnow, *Pimephales promelas*: Narcotic Industrial Chemicals. *Canadian Journal of Fisheries and Aquatic Science,* 40:743-748.

Wiener, J.G., C.C. Gilmour, and D.P. Krabbenhoft. 2003. Mercury strategy for the Bay-Delta Ecosystem: a unifying framework for science, adaptive management, and ecological restoration. Final report to the California Bay Delta Authority for Contract 46000001642. http://calwater.ca.gov/Programs/Science/adobe_pdf/MercuryStrategy_FinalReport_1-12-04.pdf

Word, J.Q., W. Gardiner, and D.M. Moore. In Press. The Influence of Confounding Factors on SQGs and Their Application to Estuarine and Marine Sediment Evaluations, Chapter 15, *In:* (R. Wenning, ed.) Use of Sediment Quality Guidelines and Related Tools for the Assessment of Contaminated Sediments.