Evaluation and Use of Sediment Reference Sites and Toxicity Tests in San Francisco Bay

Final Report

April 1998

California State Water Resources Control Board

San Francisco Bay Regional Water Quality Control Board

California Department of Fish and Game
Marine Pollution Studies Laboratory

Institute of Marine Sciences
University of California, Santa Cruz

Evaluation and Use of Sediment Reference Sites and Toxicity Tests in San Francisco Bay

Final Report

April 1998

California State Water Resources Control Board

San Francisco Bay Regional Water Quality Control Board

California Department of Fish and Game
Marine Pollution Studies Laboratory

Institute of Marine Sciences
University of California, Santa Cruz

Evaluation and Use of Sediment Reference Sites and Toxicity Tests in San Francisco Bay

April 1998

Authors

John W. Hunt, Brian S. Anderson, Bryn M. Phillips, John Newman and Ron Tjeerdema
Institute of Marine Sciences, University of California, Santa Cruz

Mark Stephenson, Max Puckett
California Department of Fish and Game, Marine Pollution Studies Laboratory

Russell Fairey
San Jose State University, Moss Landing Marine Laboratories

Robert W. Smith
EcoAnalysis

Karen Taberski
San Francisco Bay Regional Water Quality Control Board

Copies of this report may be obtained from:

Bays and Estuaries Unit, Division of Water Quality, California State Water Resources Control Board, P.O. Box 100, Sacramento, CA 95812-0100

Acknowledgments

We are grateful to the following individuals and institutions for their participation in this study:

Program Management:
Craig J. Wilson, Gita Kapahi, and Fred LaCaro
California State Water Resources Control Board

Project Direction:
Karen Taberski and Peter Otis
San Francisco Bay Regional Water Quality Control Board

Project Management and Coordination:
Max Puckett and Mark Stephenson
California Department of Fish and Game, Marine Pollution Studies Laboratory

Sample Collection:
Russell Fairey, Ross Clark, Cassandra Roberts, James Downing, Stewart Lamerdin, Eli Landreau, and Michelle Jacobi
San Jose State University, Moss Landing Marine Laboratories

Toxicity Testing:
Brian Anderson, John Hunt, Bryn Phillips, Matthew Englund, Witold Piekarski, and Michelle Hester
Institute of Marine Sciences, University of California, Santa Cruz

Trace Metal Chemical Analysis:
Mark Stephenson, Kim Paulson, Gary Ichikawa, Jon Goetzl, Jim Kanihan
California Department of Fish and Game, Marine Pollution Studies Laboratory

Trace Organic Chemical Analysis:
Ron Tjeerdema, John Newman, Matthew Stoetling, and Laura Zirelli
Institute of Marine Sciences, University of California, Santa Cruz

Total Organic Carbon Analysis:
Craig Hunter
San Jose State University, Moss Landing Marine Laboratories

Sediment Grain Size Analysis:
Pat Iampietro
San Jose State University, Moss Landing Marine Laboratories

Data Management:
Russell Fairey, San Jose State University, Moss Landing Marine Laboratories
Gary Ichikawa, Department of Fish and Game, Marine Pollution Studies Laboratory
Bryn Phillips, Shirley Tudor, and John Newman, Institute of Marine Sciences, UCSC

Statistical Design:
Robert W. Smith
EcoAnalysis, Inc.

Table of Contents

List of Figures

List of Figures (Continued)

List of Tables

List of Appendices

Executive Summary

1.0 STUDY GOALS AND OBJECTIVES

This study was undertaken as part of the California Bay Protection and Toxic Cleanup Program (BPTCP). The goal of the BPTCP is to maintain or improve environmental quality in the State's bays and estuaries by identifying and protecting relatively unpolluted areas from inflows of toxic chemicals, by identifying areas where concentrations of pollutants are associated with adverse impacts on aquatic life and/or human health, by planning for the cleanup and/or remediation of toxic sites, and by determining concentrations of chemicals in sediments that are associated with degradation of biological resources. To date, the primary focus of the program has been the identification of toxic hot spots, localized areas where elevated concentrations of toxic pollutants are found in association with adverse biological impacts.

Implicit in the definition of a toxic hot spot is the assumption that pollution in a localized area is worse than in surrounding areas, either in the same water body or in the Region where the hot spot exists. The goal of the current study was to adequately characterize ambient conditions in a water body, San Francisco Bay, to provide a standard against which to compare measurements from sites being investigated as possible hot spots. However, since program goals are to manage the State's bays and estuaries to promote environmental quality, it is not sufficient to simply characterize the "average" condition of a water body, but instead the goal of this study was to characterize the "optimal ambient conditions" currently existing. Therefore, this study focused on the identification and evaluation of sediment reference sites, the least polluted fine-grained sediment sites that could be found in San Francisco Bay with reasonable sampling effort. Reference site evaluations were based on criteria established by reviewing relevant scientific literature and consulting with the BPTCP Scientific Planning and Review Committee.

To meet this goal and to support continuing BPTCP investigations, this study focused on four objectives:

1) to identify and evaluate sediment reference sites in San Francisco Bay,
2) to evaluate appropriate sediment toxicity test methods for use in San Francisco Bay,
3) to evaluate a statistical method (the "reference envelope approach") that uses toxicity test data from reference sites to establish relative standards against which to compare results from test sites, and
4) to investigate the use of toxicity identification evaluations (TIEs) in determining the causes of toxicity at sites with both high and low concentrations of measured pollutants. Results of investigations to address this fourth objective were presented in a previous report.

2.0 TASKS ACCOMPLISHED

Seven sites were selected as candidate reference sites based on available data. Criteria for acceptable sediment reference sites included low levels of toxic chemicals, sediment grain size profiles similar to depositional areas that often serve as sinks for anthropogenic chemicals, and location remote from pollution sources. A number of reference sites were investigated to encompass the major reaches of the Bay and to cover a wide salinity range. Benthic ecological criteria were secondary in reference site selection, though some investigations of benthic community structure were undertaken through cooperation with the San Francisco Bay/Delta Regional Monitoring Program (RMP). The condition of resident benthic biological communities is often considered a critical indicator of sediment quality, but salinity fluctuations and successive invasions of exotic species in San Francisco Bay cause a high degree of variability in species composition, making it difficult to resolve pollution impacts on benthic communities.

Two sites in San Pablo Bay (North San Francisco Bay), one site in Central San Francisco Bay, two sites in the South Bay, and two sites outside of the Bay were investigated. They were identified as Island #1, Tubbs Island, Paradise Cove, North South Bay, South South Bay, Marconi Cove (Tomales Bay) and Audubon Canyon (Bolinas Lagoon), respectively. Three stations were established at each site. Surveys were conducted during three separate seasons, late Summer 1994 and late Winter/early Spring 1994 and 1995. Three stations at each of the three sites in the North and Central Bay were sampled during each survey (27 samples), while sites outside the Bay (Tomales Bay and Bolinas Lagoon) were sampled less frequently as sampling effort was directed toward the two sites in the South Bay. A total of 43 reference site samples were collected for the analyses described below. In addition, three potentially polluted sites were sampled once each for comparison.

Sediment grain size and total organic carbon (TOC) content were measured in all samples. All samples were tested for sediment toxicity, using up to nine different toxicity tests per sample. A series of standard toxicity tests was conducted on every sample, including tests of homogenized sediment using the amphipods *Ampelisca abdita* and *Eohaustorius estuarius*, and tests of sediment porewater using embryos of two invertebrates, the bay mussel *Mytilus sp.* and the sea urchin *Strongylocentrotus purpuratus*. Additional sediment toxicity tests were conducted on substantial subsets of reference site samples. These additional tests were designed to address specific study needs, such as screening for TIEs or evaluating the effects of sediment homogenization, and included sediment porewater and intact sediment cores tested

with the amphipod *Eohaustorius estuarius*, homogenized sediment tests with both the polychaete worms (*Neanthes arenaceodentata*) and the Leptostracan crustacean *Nebalia pugettensis*, and tests of sea urchin embryos exposed at the sediment-water interface. Concentrations of sediment ammonia and hydrogen sulfide were measured in all toxicity tests.

Sediment chemistry was measured at all reference sites sampled during two of the three surveys. Chemical analyses included measurement of 16 trace elements, 36 pesticides, 24 PCB congeners, and 24 polycyclic aromatic hydrocarbons (PAHs).

The two San Pablo Bay sites and the Central Bay site were sampled for benthic community analysis three times as part of the RMP pilot study. These samples were evaluated based on the presence or absence of organisms known to be indicative of either degraded or non-degraded sediments.

Toxicological, chemical, and physical measurements were analyzed to determine significant correlations and to evaluate potential sediment toxicity test reference sites. In addition to evaluating reference sites, the data were used to evaluate the nine toxicity test protocols to determine which were most useful in the Bay. Data from all San Francisco Bay reference sites were used to establish a population of reference site toxicity values (the "reference envelope") that could be used to determine tolerance limits against which to compare the results of test sites in future sediment toxicity surveys. This evaluation of reference envelope tolerance limits included additional reference site data from BPTCP hot spot screening surveys and RMP semi-annual Bay-wide surveys.

Toxicity identification evaluations (TIEs) designed to investigate the causes of sediment toxicity were conducted at four sites: one site remote from sources of pollution (Tomales Bay), and three sites heavily influenced by human activities [Islais Creek, Mission Creek (China basin), and Guadalupe Slough]. The results of these TIE investigations are presented in a separate report (Hansen et al. 1996).

3.0 MAJOR FINDINGS

3.1 Evaluation of Reference Sites

Reference site sediments consisted primarily of fine-grained silts, clays and colloids, similar to the grain size regime characteristic of depositional areas where pollutants accumulate. Total organic carbon ranged from 0.74% to 2.39% at reference sites, a range that covered TOC values found at most test sites.

Anthropogenic chemicals were generally found at relatively low concentrations at reference sites. Two elements, nickel and chromium, derived primarily from geologic sources, were found at moderately high concentrations at all sites. Sediment quality guideline values were exceeded for three other chemicals, the PAH dibenz[a,h]anthracene in a Paradise Cove sample, and the pesticide products p'p'DDT and total DDT, which were highly elevated in a measurement from a San Pablo Bay Island #1 sample. A replicate analysis of the Island #1 sample failed to duplicate the high DDT values (they were not detected in the second replicate), and the high variability indicated that the mass of these chemicals may have been small, highly concentrated, and of uncertain biological significance (the sample was not toxic in any test).

The RMP pilot study of benthic community structure included data from three of the reference sites. Island #1 was found to have a relatively low incidence of species characteristic of non-impacted conditions and a relatively high incidence of species characteristic of impacted conditions. This site was tentatively identified as being moderately impacted by pollutants. Paradise Cove had opposite proportions of indicator species, indicating a non-impacted benthic fauna, and Tubbs Island was intermediate between the two in terms of possible pollutant impacts. These results were considered preliminary, and the variable and unstable nature of benthic communities in San Francisco Bay increases the uncertainty inherent in these characterizations. But the combined ecological and chemical data indicate that these sites are clearly not pristine, but may adequately represent the least polluted sites likely to be encountered given the constraints of fine grain size, Bay-wide distribution, and logistical concerns of accessibility and sampling effort.

Toxicity tests of sediments from reference sites in San Francisco Bay produced rates of survival, growth and normal larval development that were similar to those observed in laboratory controls. Results from the two standard embryo/larval development tests (using mussels and sea urchins) in porewater were always greater than 85% of control values, as were results of sea urchin embryo/larval tests at the sediment-water interface. The two standard amphipod tests of Bay reference sediments generally produced results greater than 80% of test controls, and always greater than 60% of control values. One of 33 *Ampelisca* tests was below 80%, while 9 of 33 *Eohaustorius* tests were below 80%. Survival of polychaete worms was similar to that of *Eohaustorius* in reference site sediments, though worm growth was more variable. The three test protocols designed for specific applications (*Eohaustorius* in intact cores and porewater, and *Nebalia* in homogenized sediment) produced highly variable results, as discussed below.

3.2 Evaluation of Sediment Toxicity Tests

Criteria for evaluating sediment toxicity tests included test success rate, variability between laboratory replicates, tolerance to fine-grained sediments, tolerance to ammonia and hydrogen sulfide, salinity range, and ability to discriminate between sediments from impacted and reference sites. Tests with the amphipod *Eohaustorius* in homogenized sediment ranked well in all categories. The test is tolerant of a wide range of grain sizes and salinity, is tolerant to moderate concentrations of ammonia and hydrogen sulfide, met all test acceptability criteria, and distinguished between sediments from reference and impacted sites. Tests with the amphipod *Ampelisca* had lower salinity tolerance, and met test acceptability criteria only when test organisms from the east coast were used. However, it was similar to *Eohaustorius* in its sensitivity to test sediments and in its tolerance of ammonia and hydrogen sulfide, and it may have greater tolerance to fine-grained sediments. Embryo/larval tests with sea urchins met all test acceptability criteria and are sensitive to pollutants, but their salinity tolerance is limited. Embryo/larval tests with mussels have greater salinity tolerance and comparable toxicant sensitivity, but had a lower test success rate. Both of these tests are sensitive to ammonia and hydrogen sulfide, which limits their applicability in porewater exposures.

An additional consideration in estuarine porewater testing is that sample salinity adjustment causes variable dilution of other sample constituents, including pollutants (see Methods Section 3.2). The amount of sample dilution is dependent on the original salinity, and thus samples from the same survey may be tested at varying levels of dilution, complicating comparisons of test results among sites. Sediment/water interface (SWI) exposure systems minimize problems associated with salinity, variable sample dilution, ammonia, and hydrogen sulfide, while increasing the ecological relevance of these embryo/larval tests. In SWI exposures, embryos are held on a screen one centimeter above the sediment surface in clean overlying water. Overlying water from the same source is used for all samples, so that the effects observed are only those caused by constituents fluxing from the test sediments. Embryos may be similarly exposed in natural settings when they settle to the sediment to develop before hatching.

The polychaete worm *Neanthes* and the Leptostracan crustacean *Nebalia* are both very tolerant of ammonia and hydrogen sulfide, but were also unaffected by sediments from highly impacted sites. *Neanthes* tests met control acceptability criteria in two of three trials, while *Nebalia* tests, which were under development as a sulfide tolerant method, failed in two of three trials. Tests of amphipods in both intact cores and porewater produced highly variable results. These tests

were employed for specific purposes, and are not likely to be recommended for monitoring surveys without extensive modification.

3.3 Evaluation of Reference Envelope Statistical Method

The "Reference Envelope" approach was developed to provide an appropriate statistical method for determining whether conditions at test sites were significantly worse than those in the surrounding area. This objective is different from that of determining absolute sample toxicity. Rather than comparing results of test samples with laboratory controls using laboratory replicate variance as the statistical test variance component, the reference envelope method establishes tolerance limits based on test results from reference site samples. Tolerance limits are calculated to identify samples significantly more toxic than a chosen proportion of the reference site distribution, and statistical significance is determined using variation among reference site results. In this way, the method considers all relevant sources of variation that could affect comparisons between sites, such as variation in time and space, the interaction of time and space components, and variation between replicates (the error term). If natural factors such as grain size vary among reference sites or between surveys, then the effects of these factors are accounted for in the analysis. Any additional variation (i.e. increased toxicity) is assumed to be the result of increased pollution at test sites.

Reference site data from this study, from BPTCP hot spot screening surveys, and from RMP Bay-wide surveys were used in the calculation and evaluation of tolerance limits. All toxicity test protocols produced data that were normally distributed. Of 238 reference site values, eight were identified as outliers, using a conservative statistical outlier detection method. Tolerance limits calculated from this data set varied with data distribution, occurrence of outliers, reference envelope "p" values, and method of calculation. The "p" value is the proportion of the reference site distribution selected for the tolerance limit. For example, a "p" value of 10 would set the tolerance limit such that any sample with a test result below the limit would be as toxic or more toxic than the worst 10% of samples expected in the water body characterized by the reference sites.

Tolerance limits were highest when calculated from data with high mean values and low variability among reference sites. The sea urchin embryo/larval development test had the highest tolerance limits (e.g. 93% of the control value at a "p" value of 10). Such high tolerance limits are indicative of consistently high reference site values, but do not necessarily indicate that the level of response was biologically significant. In such cases, we would recommend deferring to a "detectable difference" criterion (such as described by Thursby et al.,

1997). Data sets with relatively low values and high variability often produced tolerance limits that were negative. Toxicity test standards below zero clearly have no utility, and these data cannot be used in this approach. The amphipod tests using homogenized sediment had tolerance limits ranging from 60% to 70% of control values (for p = 10).

An additional element of the study involved the selection of appropriate methods for calculating tolerance limits. Three methods were evaluated. One method was non-parametric, the second used a "naive" variance that assumed only a single source of variation (such as when all sites are sampled at one time, or only one site is sampled often), and the third method assumed multiple sources of variation, which is appropriate for this and most other studies, but involves more elaborate calculations. The single and multiple variation methods produced similar results when most of the variance in the data set was distributed in the error variance component. When variance was distributed more evenly among time, space, interaction, and error components, the results of the two methods diverged. Non-parametric tolerance limits depended on the absolute range of toxicity values in the reference site data set.

Appropriate application of the reference envelope approach and the resulting tolerance limits will depend on professional judgment in determining the quality of the reference data base, selection of "p" values, and suitability to the goals of the investigation. Reference site data bases with less than about six values probably cannot produce acceptable tolerance limits, and tolerance limits based on less than twenty reference site values should be applied with caution. This method can effectively distinguish impacted sites from optimal ambient conditions if those conditions are well characterized and the assumptions of the method are met. In some cases, entire water bodies may be polluted to the extent that optimal ambient conditions are not a sufficient standard for comparison, and other methods would need to be applied to measure and improve environmental quality.

Results of this study indicate that the reference sites evaluated are not pristine, but have relatively low concentrations of pollutants, and probably approximate optimal ambient conditions for fine-grained sediments in San Francisco Bay. Many toxicity test protocols produced distributions of reference site data that could be used to calculate reasonable toxicity tolerance limits. Successful application of this information for monitoring activities will require continued sampling of reference sites coincident with monitoring surveys, and thoughtful selection of reference envelope "p" values, based on careful consideration of data quality and study objectives.

Introduction

1.0 BACKGROUND

San Francisco Bay is typical of estuaries worldwide in that it provides critical habitat for aquatic species, including many commercially and ecologically important marine species that use estuaries as rearing grounds for early life stages (Conomos et al., 1979). It is also typical in that it supports tremendous economic and industrial activity related to its international port facilities that take advantage of the Bay's natural harbor. Industry, population growth, and pesticide applications over the vast agricultural area that drains to the Bay have resulted in historical and current inflows of toxic chemicals (SFEP/AHI, 1991). Public concern for human health, aquatic life and other beneficial uses of Bay waters has prompted continuing efforts to understand and monitor the effects of pollutants. The goal of the present study is to assist in the assessment of pollution impacts by evaluating methods to determine whether adverse biological responses observed in samples from San Francisco Bay are caused by localized concentrations of pollutants in Bay sediments or by factors operating on a wider scale throughout the Bay. This is part of an effort by the Regional Water Quality Control Board and the State Water Resources Control Board to identify toxic hot spots in California's bays and estuaries through the Bay Protection and Toxic Cleanup Program. Determination of statistically significant toxicity relative to responses observed at reference sites representing optimal ambient conditions found within the Bay will help to identify and prioritize sites for regulatory and/or remedial action.

Chemical pollutants entering aquatic environments commonly bind to particulate matter and tend to accumulate in sediments. The fate of pollutants in sediments is regulated by complex geochemical processes that control the availability of these chemicals to infaunal and water column biota. Because chemical bioavailability in sediments is difficult to predict, and because varying concentrations of numerous anthropogenic chemicals have been measured in sediments from locations throughout the Bay, it is difficult to determine whether chemicals found at study sites are likely to result in adverse impacts to biological communities. Knowledge of sediment chemical concentrations alone is currently insufficient to accurately predict biological effects on a site-specific basis (Long et al., 1998), and most investigations include effects-based measurements using biological indicators. A weight-of-evidence approach involving collection of synoptic chemistry, benthic ecology, and toxicity data is particularly useful in determining the probability of biological impacts from polluted sediments (Chapman et al., 1987).

While pollution effects may occur at various levels of biological organization from enzymes to ecosystems, the current study has focused on impacts to individual organisms, as measured in sediment toxicity tests (bioassays). Toxicity tests, measuring survival, growth, and normal development of aquatic organisms after laboratory exposures to sediment samples, are commonly used in regulatory assessments. Sediment toxicity tests alone do not provide sufficient information to allow an understanding of processes controlling the biological impacts of pollutants, and have a limited ability to predict damage to natural ecosystems (Luoma and Carter, 1993). However, they are useful tools for identifying toxic sediments for a number of reasons. Sediment toxicity tests can be simple, of short duration, and precise for statistical analyses (Swartz et al., 1985a). The test organisms exhibit quantifiable, obviously detrimental responses to the integrated effects of sediment contaminants, and relationships between toxicity tests results and benthic community indices have been demonstrated along contamination gradients (Swartz et al., 1982, 1985).

2.0 STUDY OBJECTIVES

The present study had four objectives:

1) identify and evaluate sediment reference sites in San Francisco Bay,
2) evaluate appropriate sediment toxicity test methods for use in San Francisco Bay,
3) evaluate a statistical method (the "reference envelope" approach) that characterizes toxicity test responses expected from samples in the absence of severe localized pollution to provide a relative standard against which to compare results from test sites, and
4) investigate the use of toxicity identification evaluations (TIEs) in determining the causes of toxicity at sites with both high and low concentrations of measured anthropogenic chemicals.

2.1 Identification of Sediment Reference Sites

The first objective was to identify and evaluate reference sites in San Francisco Bay. Previous studies have attempted to identify and set criteria for sites from which reference sediment samples could be collected (USEPA, 1986; PTI, 1991). Criteria include low concentrations of anthropogenic chemicals, distance from known major sources of pollution, and natural features such as grain size and total organic carbon (TOC) that are similar to test sediments (PTI, 1991). Potential reference sites for this study were selected on the basis of these factors. Bulk sediment trace metal and organic chemistry analyses were conducted as part of this study and were included in the reference site selection process.

Analysis of benthic community ecology was secondary in the selection of reference sites, because benthic communities in San Francisco Bay, to a greater extent than in many other estuaries, are often dominated by introduced opportunistic species whose abundance is strongly affected by seasonal salinity fluctuations (Nichols and Thompson, 1985). While benthic ecological assessments are among the best indicators of sediment quality (Chapman et al., 1987; Swartz et al., 1985b), variability in species composition in San Francisco Bay due to salinity fluctuations and successive waves of invading species were expected to substantially limit the use of ecological data in reference site selection. However, through a cooperative effort with the San Francisco Estuary Regional Monitoring Program (RMP), benthic community data were collected at three candidate reference sites. Species assemblages at these sites were characterized and compared to those from other sites throughout the Bay/Delta.

2.2 Evaluation of Sediment Toxicity Test Methods

The second objective of this study, evaluation of appropriate sediment toxicity test methods for use in San Francisco Bay, is part of continuing efforts to select monitoring tools that effectively distinguish areas impacted by pollution. Of the numerous toxicity test species and protocols available, the constraints of salinity, grain size and seasonal factors in the Bay have limited the number of tests suitable for regulatory application (Long et al., 1990). Sediment toxicity tests were chosen for evaluation in this study based on a number of criteria, including ecological relevance, wide acceptance in the scientific community, sensitivity to pollutants, success rate, precision among replicates, and tolerance to natural factors of salinity, grain size, sulfide, and ammonia.

2.3 Evaluation of the Reference Envelope Statistical Approach

2.3.1 *Reference Envelope Concept*

The third objective of this study was to evaluate a statistical method that could identify significantly toxic sites based on comparisons with toxicity data from reference sites within San Francisco Bay. As with chemical measurements, toxicity tests yield data on a continuum from low values to high, and it is necessary to distinguish between sites where toxicity is clearly indicative of localized pollution and sites where test results are characteristic of less impacted areas of the Bay. Since samples from a group of study sites would be expected to exhibit some level of variation in toxicity test response even in the absence of pollution, a method is required to determine what level of test response is significantly greater than expected of samples representing optimal ambient conditions in the Bay. The reference envelope approach uses reference site data to calculate tolerance limits

as relative standards against which to compare results from test sites. Tolerance limits are specified by a given proportion ("p") of the reference distribution, and samples that are more toxic than the tolerance limit would be considered among the most toxic "p^{th}" percentile of the Bay as represented by the reference sites. Specific details of this approach are given in part 10.3 of the Methods section.

2.3.2 *Optimal Ambient Conditions*

The term "optimal ambient conditions" is used throughout this report to indicate the least impacted state in which fine-grained sediments are likely to be found in the different basins of San Francisco Bay. It is not intended to imply "average" conditions, since the average state of Bay sediments may be unacceptable in terms of pollutant impact. The sites evaluated in this study may not be the least polluted in San Francisco Bay; substantially greater sampling effort would be needed to make that determination. However, as part of the first objective of this study, we have sought to identify sites exhibiting less human impact and chemical contamination than had been found in previous studies of other areas in the Bay. In this context, the term "ambient" is defined as representative of conditions existing over a relatively large area. Reference sites are considered to be representative of "optimal ambient conditions," rather than "background" conditions thought to exist prior to anthropogenic influence.

In an estuary as heavily urbanized as San Francisco Bay, it is probable that all sites have detectable levels of anthropogenic chemicals and some resulting potential for causing adverse biological effects. However, logistical constraints of the BPTCP require that toxic hot spot identification efforts be focused on sites where it can be clearly demonstrated that observed toxicity is due to localized pollution rather than to conditions thought to occur in a much wider geographic area.

2.3.3 *Absolute and Relative Standards*

The objective of determining significant distinctions between test sites and reference conditions is different from that of determining the absolute toxicity of a sediment sample. For this latter purpose, statistically significant sediment toxicity is often determined through comparisons of test samples against laboratory controls using standard t-test statistics (e.g., Schimmel et al., 1994). Laboratory controls are generally samples of sediment from the site where the test organisms are collected, and are thus expected to produce minimal toxicity (e.g., less than 10% mortality). The variance component of the t-test, as commonly applied, is the variance among responses from laboratory replicate test chambers. Variation among

study sites is not considered, even though toxicity test results could vary considerably among sites even in the absence of toxic chemicals. This approach, therefore, uses variability among laboratory replicates to determine whether the difference between a test sample and a control is statistically significant. We might consider the control in this case to be an absolute standard: the response of healthy animals in their native sediment. The reference envelope approach, in contrast, uses variation among reference site test results to determine the statistical significance of differences between test site results and tolerance limits calculated from reference site data. The method provides a relative standard that incorporates all types of variance that affect differences between sites over the course of a study. Variation in space (among reference sites), time (among surveys), space/time interaction, and among replicates (the error term) are all considered in determining the significance of tolerance limits.

2.3.4 *Alternative Approaches to Use of Reference Sites*

Reference sites have been used previously as relative standards for comparison with test sites (e.g., USEPA, 1986). In the simplest case, a sample from a single reference site can be compared to a test sample using laboratory replication and a t-test. Field replicates can be incorporated into experimental designs to more accurately characterize field variance, but variation within a site on a given sampling date may not adequately represent variation occurring throughout the study area over multiple sampling times that usually characterize long-term studies. A far more comprehensive method has been developed for freshwater systems, involving the use of large numbers of extensively characterized reference sites that are classified into groups using cluster analysis and ordination. Ordination scores are then correlated with non-anthropogenic variables to generate a model of how similar sites should respond in toxicity tests. Sites producing greater toxicity than predicted by the models would be considered toxic due to anthropogenic factors (e.g., Reynoldson et al., 1995). This method, however, requires a very large number of reference sites to model multiple environmental conditions, and may be difficult to implement in a setting as complex as a large estuary. The reference envelope approach evaluated in the present study is an attempt to use toxicity data from multiple reference sites to generate a population of reference values that provide a relative standard against which to compare data from test sites.

2.3.5 *Implications of Unexplained Toxicity*

An additional consideration in evaluating the reference envelope approach is that samples from sites with low levels of measured pollutants have been shown in some cases to be significantly more toxic than laboratory controls (Long et al., 1990; USEPA, 1986).

Observed toxicity from these presumably "clean" sites is unexplained, but the implications are important for regulatory decision making. If, on the one hand, the observed toxicity is due to unmeasured pollutants, and there are many anthropogenic compounds that are not routinely measured in sediment assessments, then the site should probably be targeted for further regulatory attention and not used as a reference site. However, the toxicity may be the result of natural variation, or response to naturally occurring compounds, such as algal toxins associated with fish kills (Burkholder et al., 1992). More than 1500 halogenated chemicals of natural origin have been isolated from the environment, and many have been shown to be toxic to humans, livestock, fish, mollusks, and mosquito larvae (Gribble, 1992). In such cases, this "natural" variation in toxic response should be included in the background variance component of any statistical approach used to evaluate the significance of test site data. To our knowledge, however, there is no evidence that natural toxins are responsible for observed responses in sediment toxicity tests. Care must be taken, therefore, in selecting reference sites that are indicative of ambient variability without incorporating sites with severe toxicity that may be related to unmeasured or poorly understood pollutants. A component of this study, part of the fourth objective described below, was to use toxicity identification evaluation (TIE) techniques in an attempt to understand the causes of toxicity at such sites.

2.4 Effects of Sediment Grain Size

Of the natural factors that may affect the results of toxicity tests using infaunal organisms, grain size was selected for additional analysis as part of this study. High proportions of fine-grained sediment have been shown to adversely affect test amphipods to some degree (USEPA, 1993; DeWitt et al., 1988). Oakden et al. (1984) found amphipods were capable of distinguishing between paired sediment samples with slight differences in mean grain size, and niche diversity of amphipods has been related to very specific grain size requirements (Oliver et al., 1982; Oakden, 1984; Bousfield, 1970; Dennel, 1933; Sameoto, 1969; Biernbaum, 1979; Finchham, 1969). Ott (1986) concluded that fine-grained sediments are very diverse in characteristics; some of the finest-grained sediments have less impact on amphipods if they are incorporated in organic matrices. Johnson (1974) determined that in some samples, up to 70% of the mineral grains were found in organic matter aggregates such as fecal material, and that nearly all clay and silt-size particles were incorporated into an organic matrix. Information on organic matrices is not available from standard grain size analyses. However, in one sampling period of this study, we employed microscopic analysis as proposed by Johnson (1974), in addition to the more common hydrometric techniques, to

further investigate the relationship between grain size and toxicity test response at sites with relatively low contaminant concentrations.

2.5 Toxicity Identification Evaluations

The fourth objective of this study was to investigate the use of toxicity identification evaluations (TIEs) in determining the causes of toxicity at both polluted sites and at potential reference sites that had previously produced toxic samples despite relatively low concentrations of measured contaminants. Existing TIE methods (Burgess et al., 1996; Mount, 1988; Mount and Anderson-Carnahan, 1988a, b) were evaluated, and modifications were made where necessary to adapt these methods for use with the test organisms used in this phase of the study (embryos of the purple sea urchin *Strongylocentrotus purpuratus*, embryos of bay mussels *Mytilus spp.*, and adult amphipods *Eohaustorius estuarius*). Toxic samples from three sites previously shown to be contaminated (Islais Creek, Guadalupe Slough, and East China Basin) and one site from a remote unpolluted area (Marconi Cove in Tomales Bay; Flegal et al., 1994) were chemically manipulated in attempts to selectively remove sample toxicity. By comparing the toxicity of sample fractions to the toxicity of the original sample, classes of compounds could be systematically eliminated as candidate chemicals likely to be responsible for observed toxicity at the test sites.

3.0 SAMPLING APPROACH

To accomplish the goals of this study, we conducted as many as nine different sediment toxicity tests at seven field-replicated sites in or near San Francisco Bay during three seasons. Toxicity test results were compared with sediment grain size and measured concentrations of trace metals, trace organics, ammonia, hydrogen sulfide, and total organic carbon (TOC) to investigate causes of variation in test response at selected reference sites. Data were also collected from test sites to allow an evaluation of each toxicity test's ability to distinguish between reference and impacted sites, and to determine how results from impacted sites compared with tolerance limits established using the reference envelope statistical approach. The results of these investigations are presented in this report. The TIE investigations were conducted at another facility, and are presented in a separate document (Hansen and Associates, 1996).

Methods

1.0 SITE SELECTION

1.1 Reference Sites

Seven sites were selected for evaluation as reference sites to be used in future toxicity assessments in San Francisco Bay. Sites were evaluated based on criteria established in previous studies (USEPA, 1986; PTI, 1991; Long et al., 1990), including: low concentrations of anthropogenic chemicals, distance from known major sources of pollution, and natural features such as grain size and total organic carbon (TOC) that are similar to test sediments. Sites with fine-grained sediment were selected because most heavily polluted test sites have been found in depositional areas with fine sediments. Three field replicates were collected at each site (Appendix 1). Sites selected for initial evaluation as reference sites are shown in Figures 1 through 4 and are listed below (Table 1).

Table 1a. Reference sites evaluated and used in the development of toxicity tolerance limits.

Water Body	Location	Station #	Latitude	Longitude	Sampling Dates
Central SF Bay	Paradise Cove	20005	37,53,95N	122,27,86W	4/94, 9/94, 3/95
San Pablo Bay	Tubbs Island	20006	38,06,87N	122,25,16W	4/94, 9/94, 4/95
San Pablo Bay	Island #1	20007	38,06,72N	122,19,71W	4/94, 9/94, 4/95
South SF Bay	North Site	20013	37,34,23N	122,08,98W	3/95
South SF Bay	South Site	20014	37,32,18N	122,07,16W	3/95

Table 1b. Reference sites evaluated but not used in the development of toxicity tolerance limits.

Water Body	Location	Station #	Latitude	Longitude	Sampling Dates
Bolinas Lagoon	Audubon Cyn	20008	37,55,41N	122,40,57W	4/94
Tomales Bay	Marconi Cove	20009	38,08,36N	122,52,46W	4/94, 9/94, 3/95

1.2 Test Sites

In addition to potential reference sites, one sample was collected from each of three sites where previous studies had shown either high toxicity or high levels of toxic chemicals (e.g. Flegal et al., 1994; Anderson et al., 1995; Long et al. 1988). Data from these sites were compared against reference sites as part of the evaluation of toxicity tests and the reference envelope statistical approach. Locations of these test sites are shown in Figures 2 and 3, and are listed below.

Figure 1. Location of Study Area.

Figure 2. Overview Map of Study Sites in and near San Francisco Bay. Black stars indicate reference sites; black squares indicate test sites used for comparison. Gray stars indicate sites outside of San Francisco Bay that were investigated as potential reference sites.

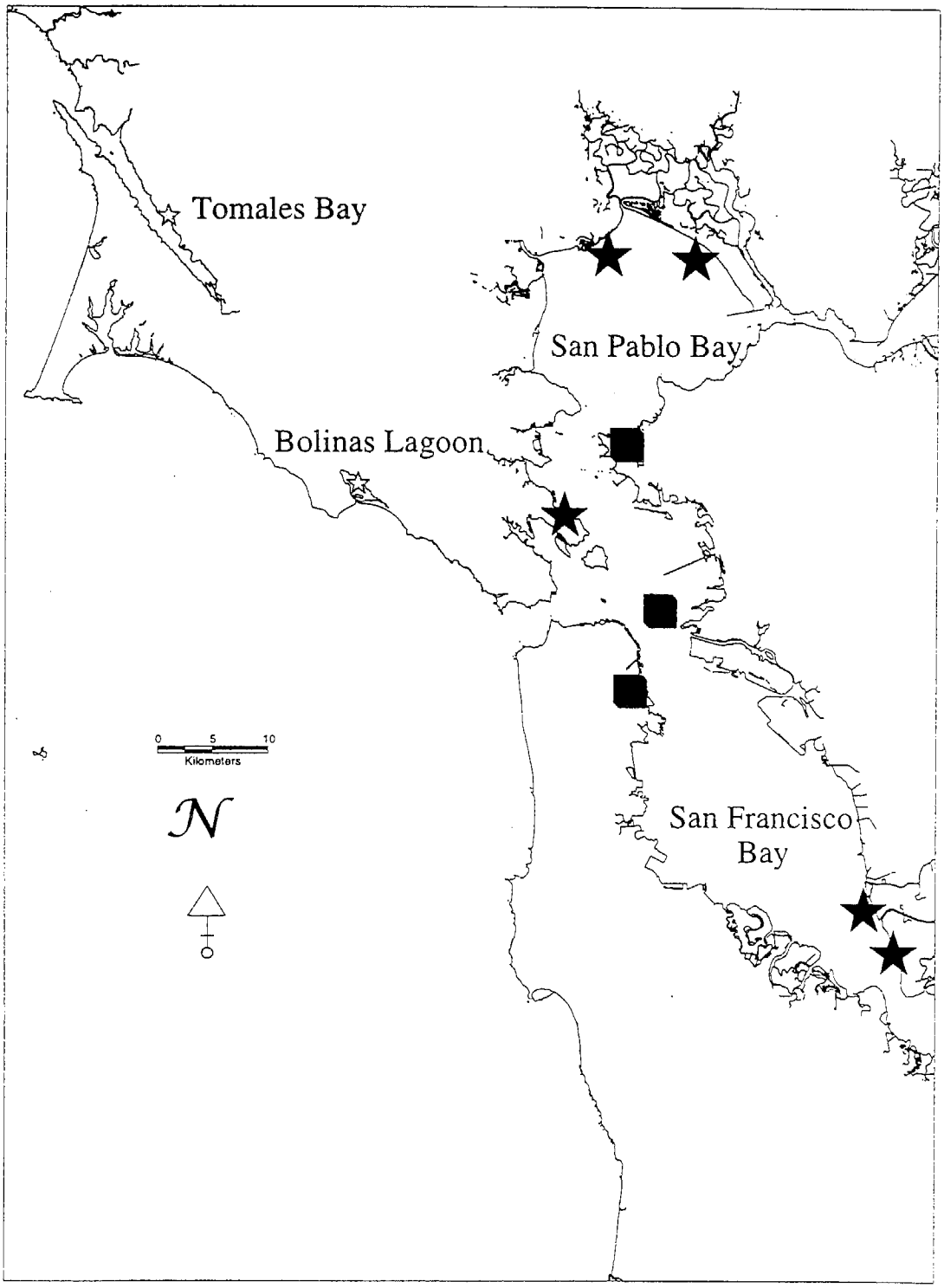


Figure 3. Location of Reference Sites in San Pablo Bay and San Francisco Bay.

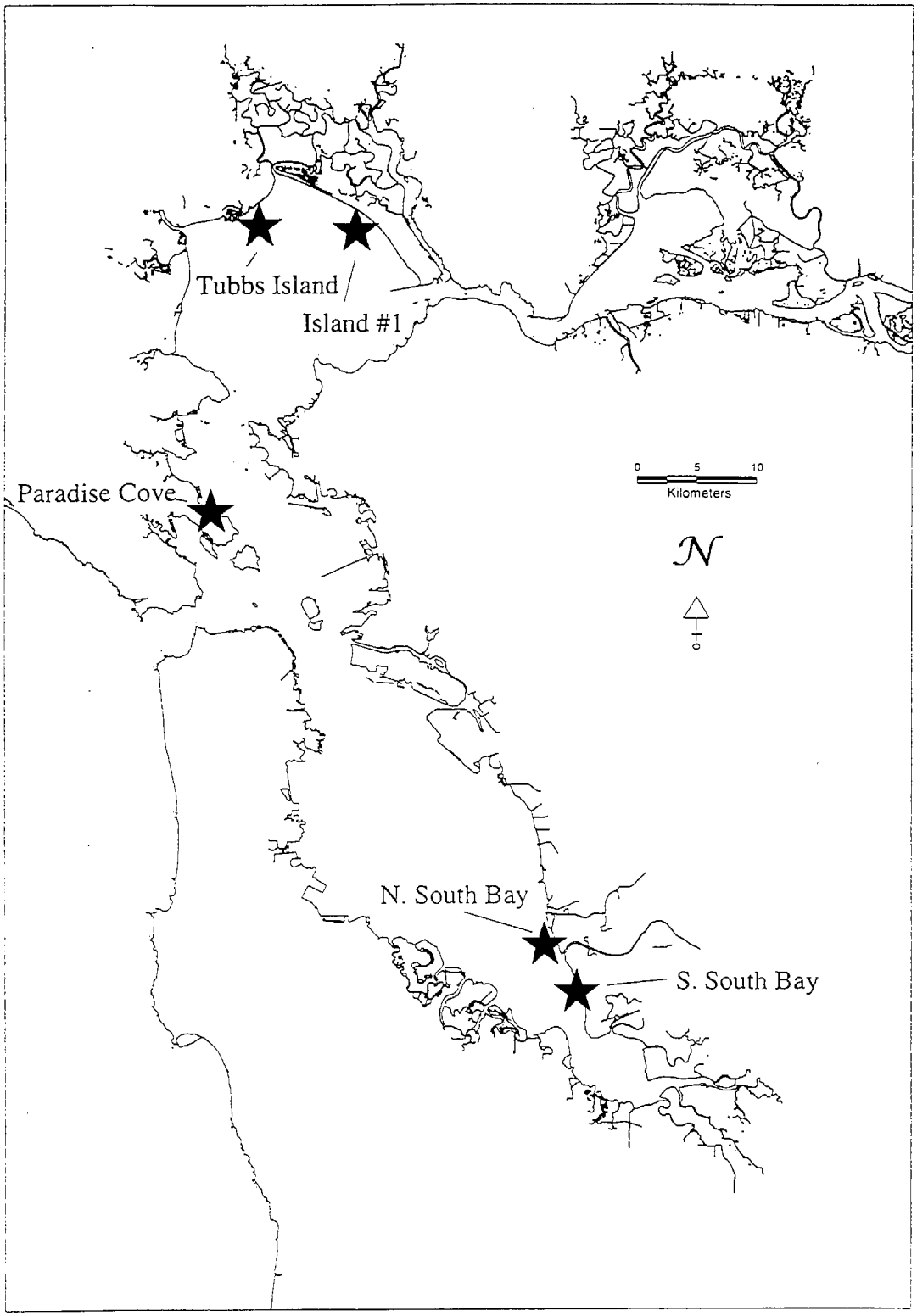


Figure 4. Location of Test Sites in San Pablo Bay and San Francisco Bay.

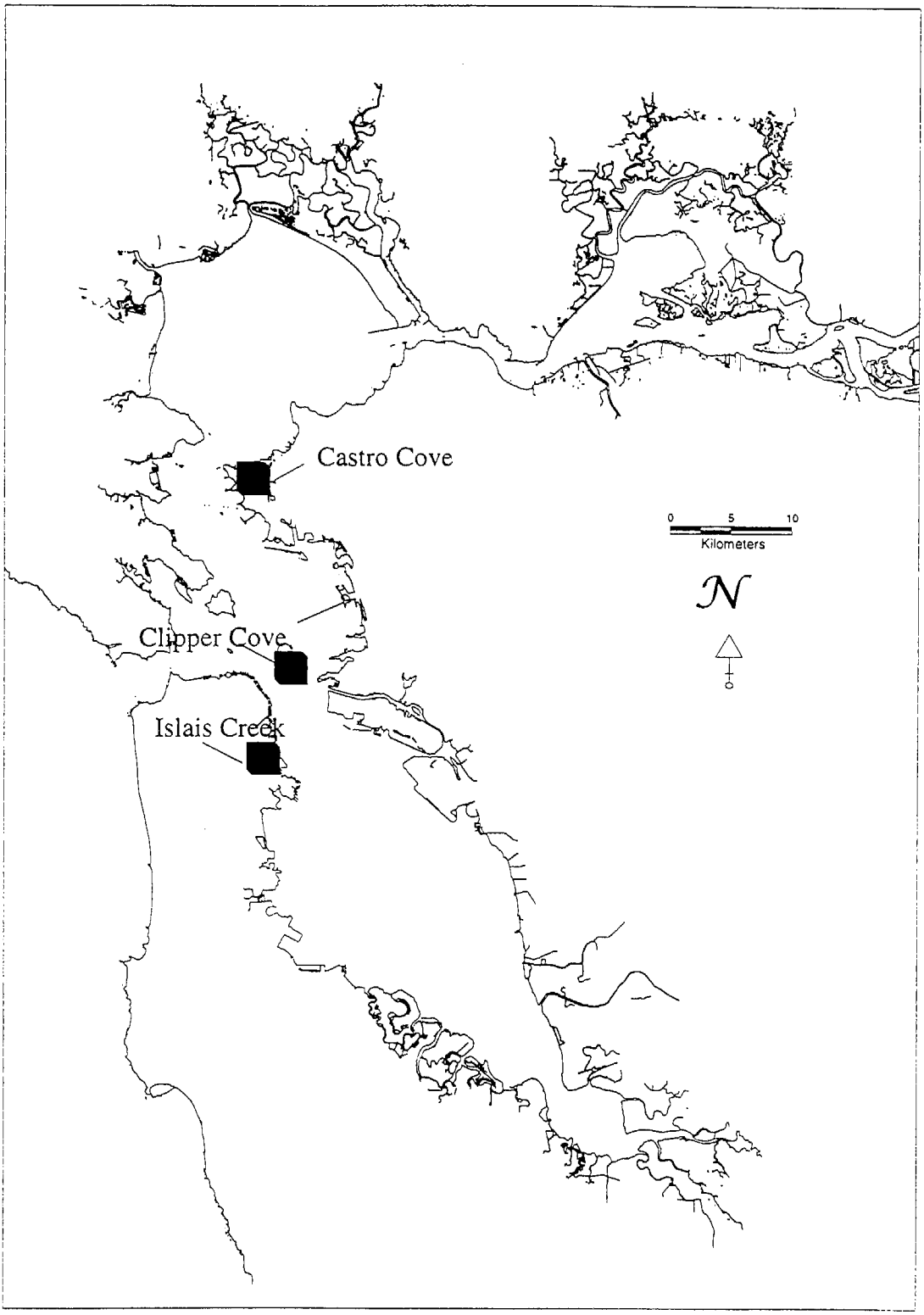


Table 2. Test Site used in evaluations of toxicity test protocols and toxicity tolerance limits.

Water Body	Location	Station #	Latitude	Longitude	Sampling Dates
San Pablo Bay	Castro Cove	20010	37,57,26N	122,24,09W	9/94
South SF Bay	Islais Creek	20011	37,44,90N	122,23,51W	9/94
Central SF Bay	Treasure Is. Clipper Cove	20012	37,48,86N	122,21,86W	3/95

2.0 SAMPLE COLLECTION AND PROCESSING

2.1 Summary Of Methods

This section describes specific techniques for collecting and processing samples. Because collection of sediments influences the results of all subsequent laboratory and data analyses, it was important that samples be collected in a consistent and conventionally acceptable manner. Field and laboratory technicians were trained to conduct a wide variety of activities using standardized protocols to ensure comparability in sample collection among crews and across geographic areas.

2.2 Cleaning Procedures

All sampling equipment (*i.e.*, containers, container liners, scoops, water collection bottles) was made from non-contaminating materials and was precleaned and packaged protectively prior to entering the field. Sample collection gear and samples were handled only by personnel wearing non-contaminating polyethylene gloves. All sample collection equipment (excluding the sediment sampler) was cleaned by using the following sequential process:

1) two-day soak and wash in Micro (brand) detergent,
2) three tap-water rinses,
3) three deionized water rinses,
4) a three-day soak in 10% HCl or HNO_3,
5) three ASTM Type II--Milli-Q -- water rinses,
6) air dry,
7) three petroleum ether rinses, and
8) air dry.

All cleaning after the Micro (brand) detergent step was performed in a positive pressure "clean" room to prevent airborne contaminants from contacting sample collection equipment. Air supplied to the clean room was filtered.

The sediment sampler was cleaned prior to entering the field, and between sampling stations, by utilizing the following sequential steps: a vigorous Micro (brand) detergent wash and scrub, a sea-water rinse, a 10% HCl rinse, and a methanol rinse. The sediment sampler was scrubbed with seawater between successive deployments at the same station to remove adhering sediments from contact surfaces possibly originating below the sampled layer.

Sample storage containers were cleaned in accordance with the type of analysis to be performed. All containers were cleaned in a positive pressure "clean" room with filtered air to prevent airborne contaminants from contacting sample storage containers.

Plastic containers (HDPE or TFE) for grain size or trace metal analysis media (sediment, archive sediment, pore water, and subsurface water) were cleaned by: a two-day Micro (brand) detergent soak, three tap-water rinses, three deionized water rinses, a three-day soak in 10% HCl or HNO_3, three Type II-Milli-Q (brand)-water rinses, and air dry.

Glass containers for total organic carbon or synthetic organic analysis media (sediment, archive sediment, pore water, and subsurface water) and additional teflon sheeting cap-liners were cleaned by: a two-day Micro (brand) detergent soak, three tap-water rinses, three deionized water rinses, a three-day soak in 10% HCl or HNO_3, three Type II-Milli-Q (brand)-water rinses, air dry, three petroleum ether rinses, and air dry.

2.3 Sediment Sample Collection

All sampling locations (latitude & longitude), whether altered in the field or predetermined, were verified using a Magellan GPS NAV 5000, and recorded in the field logbook.

The primary method of sediment collection was by use of a $0.1 m^2$ Young-modified Van Veen grab aboard a sampling vessel. Modifications included a non-contaminating Kynar coating which covered the grab's sample box and jaws. After the filled grab sampler was secured on the boat gunnel, the sediment sample was inspected carefully. The following acceptability criteria were met prior to taking sediment samples:

1) Sampler was not over-filled (*i.e.*, the sediment surface was not pressed against the top of the sampler).
2) Overlying water was present, indicating minimal leakage.
3) Overlying water was not excessively turbid, indicating minimal sample disturbance.
4) Sediment surface was relatively flat, indicating minimal sample disturbance.
5) Desired penetration depth was achieved (*i.e.*, > 5 cm).

6) Sample was muddy (>30% fines), not sandy or gravelly.
7) Sample did not include excessive shell, organic or man-made debris.
8) There were no obstructions holding the jaws open to allow sample to wash out.
If a sample did not meet all the above criteria, it was rejected.

It was critical that sample contamination be avoided during sample collection. All sampling equipment (*i.e.*, siphon hoses, scoops, containers) was made of non-contaminating material and was cleaned appropriately before use. Samples were not touched with un-gloved fingers. In addition, potential airborne contamination (*e.g.*, from engine exhaust, cigarette smoke) was avoided. Before sub-samples from the grab sampler were taken, the overlying water was removed by slightly opening the sampler, being careful to minimize disturbance or loss of fine-grained surficial sediment. Once the overlying water was removed, the top 5 cm of surficial sediment was sub-sampled from the grab. Subsamples were taken using a precleaned flat bottom Teflon scoop. This device allowed a relatively large sub-sample to be taken from a consistent depth. When subsampling surficial sediments, unrepresentative material (e.g., large stones or vegetative material) was removed from the sample in the field. Small rocks and other small foreign material remained in the sample. Criteria used to determine representativeness of sample material were determined by the chief scientist. Such removals were noted on the field data sheet. For the sediment sample, the top 5 cm was removed from the grab and placed in a pre-labeled polycarbonate container. Between grabs or cores, the sediment sample in the container was covered with a teflon sheet, and the container covered with a lid and kept cool. When a sufficient amount of sediment was collected, the sample was covered with a teflon sheet assuring no air bubbles. A second, larger teflon sheet was placed over the top of the container to ensure an air tight seal, and nitrogen was vented into the container to purge it of oxygen.

Water depth did not permit boat entrance to the Bolinas Lagoon sampling area, so divers sampled that site using sediment cores (diver cores). Cores consisted of a four-inch diameter polycarbonate tube, one-foot in length, including plastic end caps to aid in transport. Divers entered the study site from one end and sampled in one direction so as not to disturb the sediment with feet or fins. Cores were taken to a depth of at least 15 cm. Sediment was extruded out of the top end of the core to the prescribed depth of 5 cm, removed with a polycarbonate spatula and deposited into a cleaned polycarbonate tub. Additional samples were taken with the same seawater rinsed core tube until the required volume was attained. Diver core samples were treated the same as grab samples, with teflon sheets covering the sample and nitrogen purging. All sample acceptability criteria were met as with the grab sampler.

2.4 Transport Of Samples

Forty sample containers (5-liter) were packed with enough ice to keep them cool (4°± 3°C) for 48 hours. Each container was sealed in two precleaned, large plastic bags closed with a cable tie to prevent contact with other samples or ice or water. Samples were driven back to the laboratory by the sampling crew within 24 hours of collection.

2.5 Homogenization And Aliquoting Of Samples

Samples remained iced until the containers were brought back to the laboratory for homogenization. All sample identification information (station numbers, etc.) was recorded on Chain of Custody (COC) and Chain of Record (COR) forms prior to homogenizing and aliquoting. A single container was placed on plastic sheeting while also remaining in original plastic bags. The sample was stirred with a polycarbonate stirring rod for at least 5 minutes, until mud appeared homogeneous.

All prelabeled jars were filled using a clean teflon or polycarbonate scoop and stored in freezer/refrigerator (according to media/analysis) until analysis. The sediment sample was aliquoted into appropriate containers for trace metal analysis, organic analysis, pore water extraction, and toxicity testing. Samples were placed in boxes sorted by analysis type and leg number. Sample containers for sediment toxicity tests were placed in a refrigerator (4°C) while sample containers for sediment chemistry (metals, organics, TOC and grain size) were stored in a freezer (-20°C).

2.6 Procedures For The Extraction Of Pore Water

Samples were centrifuged for extraction of pore water using a Beckman JB-6 refrigerated centrifuge. One liter centrifuge bottles were filled with homogenized sediment and balanced to a uniform weight. Four bottles were centrifuged simultaneously for 30 minutes at 4°C and 2500g (3150 RPM). Supernatant porewater was siphoned from the bottles, after centrifugation, and placed in subsample containers suitable for appropriate subsequent analysis.

2.7 Chain Of Custody And Records

Chain-of-custody documents were maintained for each station. Each form was a record of all sub-samples taken from each sample. IDORG number (a unique identification number for only that sample), Department of Fish and Game (DFG) station numbers and station names, leg number (sample collection trip batch number), and date collected were included on each sheet. A chain-of-custody form accompanied every sample so that each person releasing or receiving a subsample signed and dated the form.

2.8 Authorization/Instructions To Process Samples

Standardized forms entitled "Authorization/Instructions to Process Samples" accompanied the receipt of any samples by any participating laboratory. These forms were completed by DFG personnel, or its authorized designee, and were signed and accepted by both the DFG authorized staff and the staff accepting the samples on behalf of the particular laboratory. The forms contain all pertinent information necessary for the laboratory to process the samples, such as the exact type and number of tests to run, number of laboratory replicates, dilutions, exact eligible cost, deliverable products (including hard and soft copy specifications and formats), filenames for soft copy files, expected date of submission of deliverable products to DFG, and other information specific to the laboratory and analyses being performed.

3.0 TOXICITY TESTING

3.1 Summary Of Methods

All toxicity tests were conducted at the DFG Marine Pollution Studies Laboratory (MPSL) at Granite Canyon. Toxicity tests were conducted by personnel from the Institute of Marine Sciences, University of California, Santa Cruz. Quality assurance criteria for all toxicity tests are given in Appendix B, Section 2. Water used as dilution water and overlying water in all toxicity tests was made from filtered (1μm) natural Granite Canyon seawater mixed with distilled water or spring water to the appropriate salinity.

Nine toxicity test protocols were employed in this study, including 10-day solid-phase tests with the amphipods *Ampelisca* and *Eohaustorius* in homogenized sediment, 10-day solid-phase tests with *Eohaustorius* in intact sediment cores, 10-day tests with *Eohaustorius* in pore water, 10-day solid-phase tests with the Leptostracan crustacean *Nebalia*, 20-day solid-phase tests with the polycheate worm *Neanthes*, 48-hour porewater tests with embryos of the mussel *Mytilus*, 72-hour porewater tests with embryos of the sea urchin *Strongylocentrotus*, and 72-hour tests with embryos of the sea urchin *Strongylocentrotus* exposed at the sediment-water interface. All tests were conducted at each of the three sampling periods, except for the *Eohaustorius* test in intact cores, the *Neanthes* test, and the *Nebalia* test, which were each conducted in two of the three sampling periods. Descriptions of the test methods are given below.

3.2 Handling Of Pore Water Samples For Toxicity Testing

Solid-phase sediment samples collected in April, 1994, and March/April, 1995, were held for less than 48 hours prior to extraction of pore water. Due to logistical constraints, samples collected in September, 1995, were held for time periods ranging from four to six days prior to extraction of pore water. After extraction, pore water samples were kept at 4°± 3°C for no longer than 48 hours prior to initiating toxicity tests in the first two sampling runs (April, 1994, and September, 1994). However, pore water samples were held (4°± 3°C) for as long as 8 days in March/April, 1995, because flooding and the collapse of the Highway 1 bridge over the Carmel River limited access to the toxicity testing laboratory at that time. Prior to testing, sample temperature, pH, salinity, and dissolved oxygen were measured in all samples to verify that water quality criteria were within the limits defined for test protocol.

Pore water samples with salinities outside specified ranges for each protocol were adjusted to within the acceptable range. Salinities were increased by the addition of hypersaline brine, 60‰ to 80‰, drawn from partially frozen seawater. Sample salinities and the amount of sample dilution necessary to adjust salinity for testing are given in Table 3. Water quality parameters were measured at the beginning and end of each test. Dissolved oxygen concentrations and pH were measured using an Orion EA940 expandable ion analyzer. Salinity was measured with a temperature compensating Reichart refractometer. Sample temperature was measured with a mercury thermometer. Total ammonia concentrations were measured using an ammonium ion specific electrode (Orion model 95-12), and sulfide concentrations were measured on a spectrophotometer using the colorimetric methylene blue method (Phillips et al., 1997, adapted from Fonselius, 1985).

3.3 Handling Of Sediment Samples For Toxicity Testing

Bedded sediment samples were held at 4°C until required for testing. All solid-phase sediment tests were initiated within 14 days of the sample collection date. All sediment samples were processed according to procedures described in ASTM (1993). Water quality parameters, including ammonia and hydrogen sulfide concentrations, were measured in one replicate test container from each sample in the overlying water as described above. Ammonia, hydrogen sulfide and pH were measured in both overlying water (collected within 1 cm of the sediment) and in interstitial water extracted by centrifugation at the beginning and end of each test. Samples for ammonia and pH were held in capped containers and measured within one hour of extraction. Hydrogen sulfide samples were preserved with zinc acetate immediately after extraction. Measurements were taken at the beginning and end of all tests, and during overlying water renewals in the *Neanthes* tests.

Table 3. Sample Pore Water Salinity and Pore Water Concentration in Test Solutions. Samples with salinity beyond the range appropriate for each protocol were adjusted with hypersaline brine or distilled water. This adjustment diluted the samples, decreasing pore water concentrations in test solutions to the levels indicated. Protocol salinity range was 32 ± 2‰ for sea urchins, 28 ± 2‰ for mussels, and 28 ± 3‰ for the amphipod *Eohaustorius estuarius*.

				% Pore Water in Test Solution after Salinity Adjustment		
Location	Station #	Test Date	Sample Salinity ‰	Sea Urchin Development (Tested at 32‰)	Mussel Development (Tested at 28‰)	Amphipod Survival (Tested at 28‰*)
Paradise Cove	20005	4/94	26‰	83%	95%	100%
Tubbs Island	20006	4/94	18	68	77	81
Island #1	20007	4/94	18	68	77	81
Audubon Cyn	20008	4/94	31	100	92	89
Marconi Cove	20009	4/94	32	100	88	89
Paradise Cove	20005	9/94	30	93	93	100
Tubbs Island	20006	9/94	25	83	93	100
Island #1	20007	9/94	24	81	92	100
Marconi Cove	20009	9/94	34	100	82	80
Castro Cove	20010	9/94	27	87	100	100
Islais Creek	20011	9/94	30	93	93	100
Paradise Cove	20005	3/95	18	75	83	83
Tubbs Island	20006	4/95	2	55	61	83*
Island #1	20007	4/95	2	56	62	85*
Marconi Cove	20009	3/95	28	90	100	100
Treasure Is. Clipper Cove	20012	3/95	20	76	86	86
North S. Bay	20013	3/95	15	71	79	79
South S. Bay	20014	3/95	15	71	79	79

* Amphipod test solution salinity for stations 20006 and 20007 was 15‰.

In cases where sample salinity was beyond 3‰ from the test target salinity, overlying water was prepared at a salinity calculated to produce the target salinity after equilibrium was reached between overlying water and sample pore water. Sediment was not stirred with overlying water, but salinity was allowed to equilibrate through flux for 24 hours prior to introduction of test organisms. Neither ammonia nor hydrogen sulfide was adjusted prior to testing.

3.4 Tests With The Amphipod *Ampelisca abdita*

Ampelisca toxicity tests followed the ASTM (1993) standard guide for *Ampelisca abdita*. All test animals for March and September, 1994, tests were collected from San Francisco Bay by John Brezina Associates. All test animals for March, 1995, tests were obtained from East Coast Amphipod in Kingston, Rhode Island. Animals were shipped via overnight courier in one gallon polyethylene jars containing collection site sediment. If necessary, upon arrival at Granite Canyon, the amphipods were acclimated to laboratory conditions by adjusting salinity and temperature by no more than 10‰ and 2°C per day to 28‰ and 15° C. *Ampelisca* holding time at MPSL varied. For the March, 1994, test, amphipods were tested the same day they arrived at the laboratory, so that they would not build tubes from which they would need to be removed for sorting prior to inoculation. For the September, 1994, test, amphipods were held at MPSL for 48 hours prior to inoculation into the test containers. Flooding during the March, 1995, test interrupted vehicle access to the laboratory. Amphipods were received from the supplier at the residence of one of the investigators, where they were held at 16°C in aerated four-liter shipping containers with home sediment for 12 hours. The shipping containers were then carried in backpacks to MPSL, where the amphipods were adjusted to test salinity. The following day, after 24 hours holding at MPSL, they were inoculated into the test containers.

One day prior to test initiation, each sediment sample was distributed into five replicate one-liter glass beakers so that each contained 2 cm of sediment. Granite Canyon seawater, diluted with distilled water or spring water, was added to fill the container to the 700 ml mark. Overlying water salinity was either 28‰ or a salinity calculated to reach 28‰ after equilibration with the sediment sample. The test sediment and overlying water were allowed to equilibrate for 24 h, then 20 amphipods were placed into each beaker along with 28‰ sea water to fill the test containers to the one-liter line. Test chambers were then gently aerated and continuously illuminated.

Five replicates of each sample were tested for 10 days at 28‰ and 15°C. In addition, 5 replicates of a negative control were tested with each set of samples. Sediments used in negative controls were either fine-grained sediment from the *Ampelisca* collection site or medium-fine sand from the *Eohaustorius* collection site. Amphipod emergence and visible survival were recorded daily. After 10 days, samples were sieved through a 0.4 mm Nitex screen to recover the test animals, and the number of survivors was recorded for each replicate. Mean percent survival per sample was the test endpoint.

Positive control reference tests were conducted concurrently with each sediment test using cadmium chloride as a reference toxicant. In these tests, amphipod survival was recorded in three replicates of four cadmium concentrations after a 96 h water-only exposure. A dilution water control consisting of one micron filtered Granite Canyon sea water was included in each reference toxicant test.

3.5 Tests With The Amphipod *Eohaustorius estuarius*

3.5.1 *Homogenized Sediment*

The *Eohaustorius* tests followed ASTM (1993) standard guide for *Eohaustorius estuarius*. All *Eohaustorius* were obtained from Northwestern Aquatic Sciences in Yaquina Bay, Oregon. Animals were separated into groups of approximately 100 and placed in polyethylene boxes containing Yaquina Bay collection site sediment, then shipped on ice via overnight courier. Upon arrival at Granite Canyon, the amphipods were slowly acclimated to laboratory conditions by adjusting salinity and temperature by no more than 10‰ and 2°C per day to 28‰ and 15°C, except for April, 1995, tests, which were conducted at 15‰ and 15°C. Once acclimated, the animals were held for at least 48 h prior to inoculation into the test containers.

One day prior to test initiation, each sediment sample was distributed into five replicate one-liter glass beakers so that each contained 2 cm of sediment. Granite Canyon seawater, diluted with distilled water or spring water, was added to fill the container to the 700 ml mark. The test sediment and overlying water were allowed to equilibrate for 24 h, then 20 amphipods were placed into each beaker along with 28‰ sea water to fill the test containers to the one-liter line. Test chambers were then gently aerated and continuously illuminated.

Five replicates of each sample were tested for 10 days at 28‰ and 15°C. In addition, a negative control consisting of five replicates of medium-fine sand from the amphipod collection site was included with each set of samples tested. Amphipod emergence and visible survival were recorded daily. After 10 days, samples were sieved through a 0.4 mm Nitex screen to recover the test animals, and the number of survivors was recorded for each replicate.

3.5.2 *Intact Sediment Cores*

Eohaustorius tests utilizing intact sediment cores were conducted simultaneously with homogenized sediment samples. Intact cores were collected from grab samples by inserting a 7.5 cm diameter polycarbonate core tube to a depth of 10 cm (Figure 5). Core tubes were capped on both ends and transported to MPSL in coolers at 4°C. One day prior to test initiation, the space overlying the sediment was filled with 28‰ water. Test sediment and

overlying water were allowed to equilibrate for 24 h, then 20 amphipods were placed in each core tube. The remainder of the test followed the procedure used with homogenized samples. Negative controls were homogenized home sediment, the same as those used in the tests of homogenized sediment described above.

3.5.3 *Porewater*

Eohaustorius pore water tests were also conducted simultaneously with homogenized samples. Five amphipods were placed in each of five replicate loosely-covered 250 ml glass crystallizing dishes containing 50 ml porewater adjusted to 28‰. Addition of hypersaline brine or distilled water for salinity adjustment diluted pore water to concentrations ranging from 79 to 100% (Table 3). Test duration was 10 days. Fifty percent of the porewater was renewed every 96 hours. Test containers were held in darkness and were not aerated. Survival was recorded at renewals and test termination. Granite Canyon seawater adjusted to 28‰ with distilled water was distributed into 5 replicate test containers to serve as negative controls. In tests where salinity adjustment of pore water samples was necessary, brine controls were included. Brine controls contained the same proportion of hypersaline brine as was used to adjust the lowest salinity sample (i.e., the maximum brine concentration).

Positive control reference tests using cadmium chloride were conducted concurrently with each *Eohaustorius* sediment or pore water test. In these tests amphipod survival was recorded in three replicates of four cadmium concentrations after a 96-h water-only exposure. A dilution water control consisting of 1 µm-filtered Granite Canyon sea water was included in each test.

3.6 Mussel (*Mytilus edulis*) Larval Development Tests

The bay mussel (*Mytilus edulis*) larval development tests were conducted on porewater samples. Details of the test protocol are given in ASTM (1993). A brief description of the method follows.

Mussels were shipped via overnight courier from Carlsbad Aquafarms and held at Granite Canyon at ambient temperature (11-13°C) and salinity (32-34‰) until testing. A few hours before test initiation, adult mussels were transferred to 28‰ water heated to 23° to 25°C to induce spawning. Spawning adults were quickly transferred to 15°C water. Sperm and eggs were mixed in 28‰ (15°C) water to give a final sperm to egg ratio of 15 to 1. After approximately 20 minutes, fertilized eggs were rinsed on a 25 µm screen to remove excess sperm, and embryos were distributed to the test containers after approximately 90% of the embryos exhibited first cell cleavage (approximately 1 hour).

Test containers were polyethylene-capped, sea water-leached, 20 ml glass scintillation vials containing 10 ml of porewater (Hunt et al., 1998). Each test container was inoculated with approximately 250 embryos (25/ml). Porewater samples were tested at 28 ± 2‰ (15°C) . Low salinity samples were adjusted to 28‰ using hypersaline brine made from freezing seawater. High salinity samples were adjusted to 28‰ using distilled water. Addition of hypersaline brine or distilled water for salinity adjustment diluted pore water to concentrations ranging from 61 to 100% (Table 3). Negative controls consisted of one micron filtered Granite Canyon sea water adjusted to 28‰. In tests where salinity adjustment of pore water samples was necessary, brine controls were included. Brine controls contained the same proportion of hypersaline brine as was used to adjust the lowest salinity sample (i.e., the maximum brine concentration). A positive control reference test was conducted concurrently with each test using a dilution series of cadmium chloride as a reference toxicant.

After a 48 h exposure period, larvae were fixed in 5% buffered formalin. All larvae in each container were examined under an inverted light microscope at 100x to determine the proportion of normally developed larvae as described by ASTM (1993). The percentage normally developed larvae was calculated as:

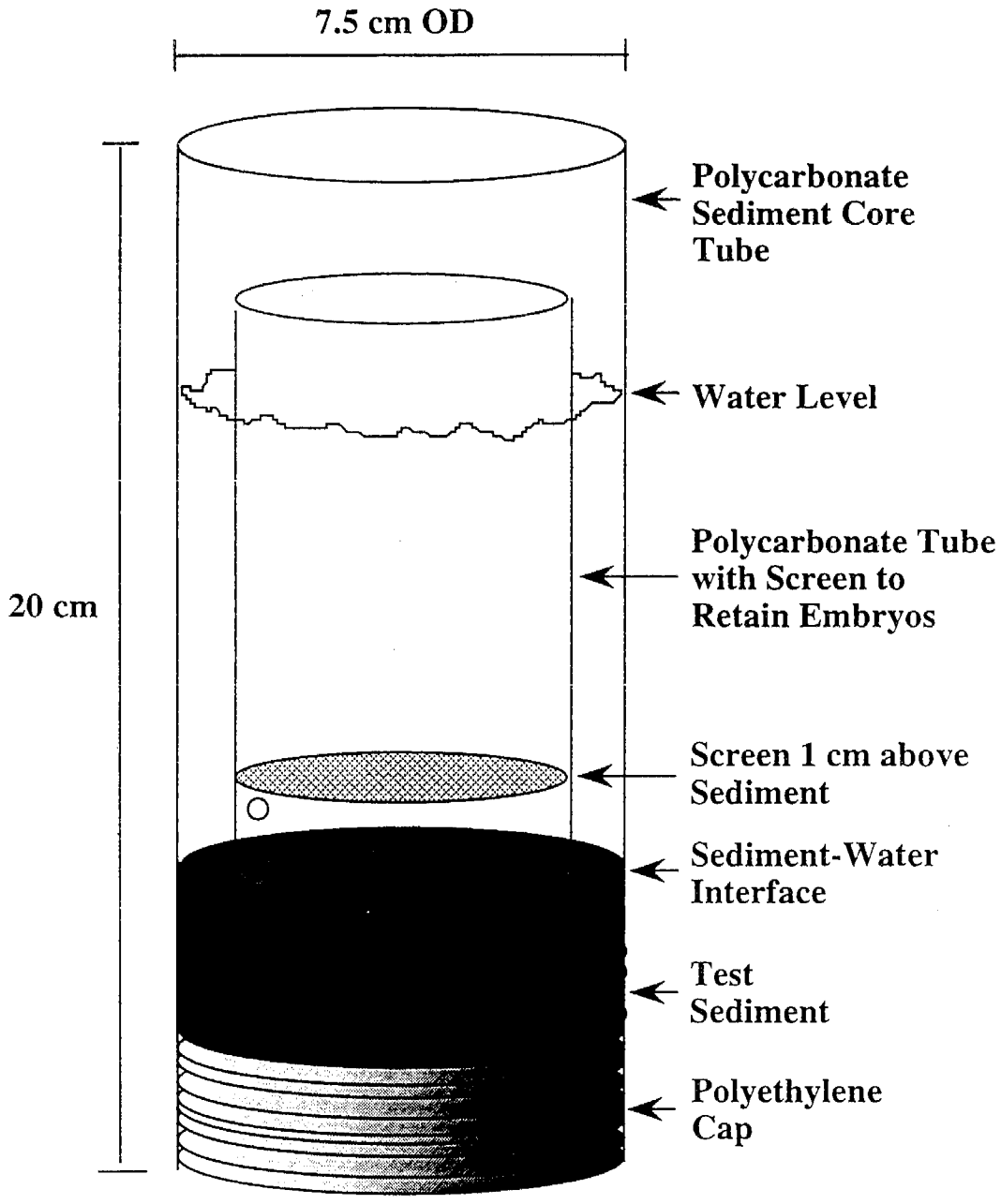
$$\frac{\text{Observed number of live normal larvae}}{\text{Mean number of live embryos inoculated at start of test}} \quad (\text{x } 100\%)$$

3.7 Sea Urchin *(Strongylocentrotus purpuratus)* Larval Development

3.7.1 *Porewater*

The purple sea urchin (*Strongylocentrotus purpuratus*) larval development test was conducted on all porewater samples. Details of the test protocol are given in Chapman et al. (1995). Sea urchins were collected from the Monterey County coast near Granite Canyon and held at ambient seawater temperature and salinity until testing. Adult sea urchins were held in complete darkness to preserve gonadal condition. On the day of a test, urchins were induced to spawn in air by injection with 0.5 ml of 0.5 M KCl. Eggs and sperm collected from the urchins were mixed in seawater at a 500 to 1 sperm to egg ratio, and embryos were distributed to the test containers within one hour of fertilization. Test containers were polyethylene-capped, seawater-leached, 20 ml glass scintillation vials containing 5 mls of porewater. Each test container was inoculated with approximately 150 embryos (30/ml). Tests were conducted at ambient seawater salinity (32 - 34‰ ± 2‰). Low salinity samples were adjusted to ambient salinity using hypersaline brine made from freezing seawater. Addition of hypersaline brine for salinity adjustment diluted pore water to concentrations ranging from 55 to 100% (Table 3).

Figure 5. Sediment-Water Interface Exposure System.
(After Anderson et al., 1996)



Laboratory controls were included with each set of samples tested. Negative controls consisted of one micron-filtered Granite Canyon sea water. In tests where salinity adjustment of pore water samples was necessary, brine controls were included. Brine controls contained the same proportion of hypersaline brine as was used to adjust the lowest salinity sample (i.e. the maximum brine concentration). A positive control reference test was conducted concurrently with each porewater test using a dilution series of copper chloride as a reference toxicant.

After an exposure period of 72 to 96 hours, larvae were fixed in 5% buffered formalin. Unpublished data has indicated no loss of test sensitivity in 72 h exposures (n = 16 reference toxicant tests compared at MPSL). One hundred larvae in each container were examined under an inverted light microscope at 100x to determine the proportion of normally developed larvae as described by Chapman et al. (1995). Percent normal development was calculated as:

$$\frac{\text{Number of normally developed larvae counted}}{\text{Total number of larvae counted}} \quad (\text{x } 100\%)$$

3.7.2 *Sediment/Water Interface*

Sea urchin larval development was also assessed at the sediment water interface (Anderson et al., 1996). This was achieved by introducing embryos into a 37 µm screen tube placed 1 cm above the sediment surface within an intact sediment core tube (Figure 5). Intact sediment cores were sampled in the same manner as was used for *Eohaustorius* amphipods tested in intact cores, described above. One day prior to test initiation, seawater at ambient salinity (33 ± 1‰) was added to fill the core tubes, and then screen tubes were added to the cores and the system was allowed to equilibrate for 24 hours. Urchin embryos were prepared as described above for pore water tests and added to the screen tubes. Each screen tube was inoculated with approximately 250 embryos. Laboratory controls consisted of Yaquina Bay, Oregon, amphipod home sediment obtained from Northwestern Aquatic Sciences.

After an exposure period of 72 to 96 hours, screen tubes were removed from the sediment cores, rinsed, and larvae were removed using a squirt bottle. Larvae were washed into 20 ml scintillation vials and fixed in 5% buffered formalin. One hundred larvae in each container were examined under an inverted light microscope at 100x to determine the proportion of normally developed larvae as described by Chapman et al. (1995). Percent normal development was calculated as:

$$\frac{\text{Number of normally developed larvae counted}}{\text{Total number of larvae counted}} \quad (\text{x } 100\%)$$

3.8 Tests With The Polycheate *Neanthes arenaceodentata*

The *Neanthes* test followed procedures described by the Puget Sound Estuary Program (PSEP, 1991). Emergent juvenile *Neanthes arenaceodentata* (2-3 week-old) were obtained from Dr. Don Reish of California State University at Long Beach, California. Worms were shipped in seawater in plastic bags at ambient temperature via overnight courier. Upon arrival at MPSL, worms were allowed to acclimate gradually to 28‰ with 2‰ daily incremental salinity adjustments at a temperature of 20° C. Once acclimated, the worms were maintained for at least 48 hours, and no longer than 10 days, before the start of a test.

The test design was similar to that described for the amphipods. One day prior to test initiation, each sediment sample was distributed into five replicate one-liter glass beakers so that each contained 2 cm of sediment. Granite Canyon seawater, diluted to 28‰ with distilled water or spring water, was added to fill the container to the 700 ml mark. The test sediment and overlying water were allowed to equilibrate for 24 h, then five worms were placed into each beaker along with 28‰ sea water to fill the test containers to the one-liter line. Test chambers were then gently aerated and continuously illuminated during the 20-day test period. Worms were fed TetraMin® every 2 days, and overlying water was renewed every 3 days.

After 20 days, samples were sieved through a 0.5 mm Nitex screen, and the number of surviving worms was recorded. Surviving worms from each replicate were wrapped in a piece of pre-weighed aluminum foil, and placed in a drying oven (60°C) until they reached constant weight (48 h). Each foil packet was then cooled in a desiccator and weighed to the nearest 0.1mg. Worm survival and mean weight/worm for each replicate was calculated as follows:

$$\text{Percent worm survival} = \frac{\text{Number of surviving worms}}{\text{Initial number of worms}} \quad (\text{x } 100\%)$$

$$\text{Mean weight/worm} = \frac{\text{Total weight - foil weight}}{\text{Initial number of worms}} \quad (\text{x } 100\%)$$

Positive control reference tests using cadmium chloride were conducted concurrently with each sediment test. In these tests, worm survival was recorded in three replicates of four cadmium concentrations after a 96-h water-only exposure. A dilution water control consisting of 1 µm-filtered Granite Canyon sea water was included in each test.

3.9 Tests With The Leptostracan Crustacean *Nebalia pugettensis*

This test has not been previously evaluated, but this organism was employed because of its potential tolerance to high levels of ammonia and hydrogen sulfide (unpublished data). Tests utilizing *Nebalia pugettensis* followed the ASTM (1993) standard guide for the marine amphipod *Rhepoxynius abronius*. Test organisms were obtained from the tidal mud flats in Elkhorn Slough near Moss Landing, California. Animals were held at ambient water temperature and salinity at the Moss Landing Marine Laboratories until test initiation. Sediment sample preparation, test initiation, and test termination were as previously described for the amphipods, except that emergence data were not collected, since the animals tend to hover at the sediment/water interface in their natural habitat. *Nebalia* tests were conducted at 15°C in 28‰ water. Sediment controls consisted of the Moss Landing beach sand that was used as a culture medium for the organisms.

Positive control reference tests were conducted concurrently with each sediment test using cadmium chloride as a reference toxicant. In these tests, survival was recorded in three replicates of four cadmium concentrations after a 96 h water-only exposure. A dilution water control consisting of one micron filtered Granite Canyon sea water was included in each test.

3.10 Toxicity Identification Evaluations (TIEs)

Toxicity identification evaluations (TIEs) were conducted using sea urchin larvae as the detector species on samples from four sites: Marconi Cove (Tomales Bay), Islais Creek, China Basin, and Guadalupe Slough. Methods for TIEs are presented in a separate report (Hansen and Associates, 1996).

4.0 TRACE METALS ANALYSIS OF SEDIMENTS

4.1 Summary Of Methods

Trace Metals analyses were conducted at the California Department of Fish and Game's Trace Metal Analytical Facility at Moss Landing, CA. Table 4 indicates the trace metals analyzed and lists method detection limits for sediments.

4.2 Analytes And Detection Limits

Table 4 - Trace Metal Detection Limits in Sediments.

Trace Element	Detection Limit (µg/g, dry weight)
Aluminum	1
Antimony	0.1
Arsenic	0.1
Cadmium	0.01
Chromium	0.1
Copper	1.0
Iron	0.1
Lead	0.1
Manganese	0.05
Mercury	0.03
Nickel	0.1
Selenium	0.2
Silver	0.01
Tin	0.02
Tributyltin	0.02
Zinc	0.05

4.3 Sediment Digestion Procedures

A one gram aliquot of sediment was placed in a pre-weighed Teflon vessel, and one ml of a concentrated 4:1 nitric:perchloric acid mixture was added. Vessels were capped and heated in a vented oven at 130° C for four hours. Three ml of hydrofluoric acid were added to the vessel, recapped and returned to the oven overnight. Twenty ml of 2.5% boric acid were added to the vessel and placed in the oven for an additional 8 hours. Weights of the Teflon vessel and solution were recorded, and the solution was poured into 30 ml polyethylene bottles.

4.4 Atomic Absorption Spectrometry Methods

Samples were analyzed by furnace AA on a Perkin-Elmer Zeeman 3030 Atomic Absorption Spectrophotometer, with an AS60 auto sampler, or a flame AA Perkin Elmer Model 2280. Samples, blanks, matrix modifiers, and standards were prepared using clean techniques inside a clean laboratory with positive pressure air filtration. ASTM Type II water and ultra clean chemicals were used for all standard preparations. All elements were analyzed with platforms for stabilization of temperatures. Matrix modifiers were used when components of the matrix interfered with adsorption. The matrix modifier was used for Sn, Sb and Pb. Continuing calibration check standards (CLC) were analyzed with each furnace sheet, and calibration curves were run with three concentrations after every 10 samples. Blanks and standard reference materials, MESS1, PACS, BCSS1 or 1646 were analyzed with each set of samples for sediments.

5.0 TRACE ORGANIC ANALYSIS OF SEDIMENTS (PCBs, PESTICIDES, AND PAHs)

5.1 Summary Of Methods

Analytical sets of 12 samples were scheduled such that extraction and analysis would occur within a 40 day window. The methods employed by the UC Santa Cruz Trace Organics Analytical Facility were modifications of those described by Sloan *et al.* (1993). Tables 5, 6 and 7 indicate the pesticides, PCBs, and PAHs analyzed and list method detection limits for sediments on a dry weight basis.

5.2 Analytes And Detection Limits

Table 5. Organochlorine Pesticides Analyzed and Their Detection Limits (ng/g dry weight) in Sediment.

Compound	Detection Limit
Aldrin	0.5
cis-Chlordane	0.5
trans-Chlordane	0.5
alpha-Chlordene	0.5
gamma-Chlordene	0.5
Chlorpyrifos	1.0
Dacthal	0.2
o,p'-DDD	1.0
p,p'-DDD	0.4
o,p'-DDE	1.0
p,p'-DDE	1.0
p,p'-DDMS	3.0
p,p'-DDMU	2.0
o,p'-DDT	1.0
p,p'-DDT	1.0
p,p'-Dichlorobenzophenone	3.0
Dieldrin	0.5
Endosulfan I	0.5
Endosulfan II	1.0
Endosulfan sulfate	2.0
Endrin	2.0
Ethion	2.0
alpha-HCH	0.2
beta-HCH	1.0
gamma-HCH	0.2
delta-HCH	0.5
Heptachlor	0.5
Heptachlor Epoxide	0.5
Hexachlorobenzene	0.2
Methoxychlor	1.5
Mirex	0.5
cis-Nonachlor	0.5
trans-Nonachlor	0.5
Oxadiazon	2.0
Oxychlordane	0.5
Toxaphene	10

Table 6. Pentachlorobiphenyls (PCB) Congeners Analyzed and Their Detection Limits* in Sediment.

NIST List of PCB Congeners	
PCB Congener 8	PCB Congener 128
PCB Congener 18	PCB Congener 138
PCB Congener 28	PCB Congener 153
PCB Congener 44	PCB Congener 170
PCB Congener 52	PCB Congener 180
PCB Congener 66	PCB Congener 187
PCB Congener 87	PCB Congener 195
PCB Congener 101	PCB Congener 206
PCB Congener 105	PCB Congener 209
PCB Congener 118	

Additional Congeners:	
PCB Congener 5	PCB Congener 137
PCB Congener 15	PCB Congener 149
PCB Congener 27	PCB Congener 151
PCB Congener 29	PCB Congener 156
PCB Congener 31	PCB Congener 157
PCB Congener 49	PCB Congener 158
PCB Congener 70	PCB Congener 174
PCB Congener 74	PCB Congener 177
PCB Congener 95	PCB Congener 183
PCB Congener 97	PCB Congener 189
PCB Congener 99	PCB Congener 194
PCB Congener 110	PCB Congener 201
PCB Congener 132	PCB Congener 203

*All individual PCB Congener detection limits were 1 ng/g dry weight.

Aroclors:	Detection Limit
Aroclor 5460	50

Table 7. Polycyclic Aromatic Hydrocarbons (PAH) Analyzed and Detection Limits in Sediment.

Compound	Detection Limit (ng/g dry weight)
Naphthalene	5
2-Methylnaphthalene	5
1-Methylnaphthalene	5
Biphenyl	5
2,6-Dimethylnaphthalene	5
Acenaphthylene	5
Acenaphthene	5
2,3,5-Trimethylnaphthalene	5

Table 7 (Continued). Polycyclic Aromatic Hydrocarbons (PAH) Analyzed and Detection Limits in Sediment .

Compound	Detection Limit (ng/g dry weight)
Fluorene	5
Phenanthrene	5
Anthracene	5
1-Methylphenanthrene	5
Fluoranthrene	5
Pyrene	5
Benz[a]anthracene	5
Chrysene	5
Benzo[b]fluoranthrene	5
Benzo[k]fluoranthrene	5
Benzo[e]pyrene	5
Benzo[a]pyrene	5
Perylene	5
Indo[1,2,3-cd]pyrene	5
Dibenz[a,h]anthracene	5
Benzo[ghi]perylene	5

5.3 Extraction And Analysis

Samples were removed from the freezer and allowed to thaw. A 10 gram sample of sediment was removed for chemical analysis and an independent 10 gram aliquot was removed for dry weight determinations. The dry weight sample was placed into a pre-weighed aluminum pan and dried at 110°C for 24 hours. The dried sample was reweighed to determine the sample's percent moisture. The analytical sample was extracted 3 times with methylene chloride in a 250-mL amber Boston round bottle on a modified rock tumbler. Prior to rolling, sodium sulfate, copper, and extraction surrogates were added to the bottle. Sodium sulfate dehydrates the sample allowing for efficient sediment extraction. Copper, which was activated with hydrochloric acid, complexes free sulfur in the sediment.

After combining the three extraction aliquots, the extract was divided into two portions, one for chlorinated hydrocarbon (CH) analysis and the other for polycyclic aromatic hydrocarbon (PAH) analysis.

The CH portion was eluted through a silica/alumina column, separating the analytes into two fractions. Fraction 1 (F1) was eluted with 1% methylene chloride in pentane and contains > 90% of p,p'-DDE and < 10% of p,p'-DDT. Fraction 2 (F2) analytes were eluted with 100% methylene chloride. The two fractions were exchanged into hexane and concentrated to 500 µL

using a combination of rotary evaporation, controlled boiling on tube heaters, and dry nitrogen blow downs.

F1 and F2 fractions were analyzed on Hewlett-Packard 5890 Series gas chromatographs utilizing capillary columns and electron capture detection (GC/ECD). A single 2 µl splitless injection was directed onto two 60m x 0.25mm i.d. columns of different polarity (DB-17 & DB-5; J&W Scientific) using a glass Y-splitter to provide a two dimensional confirmation of each analyte. The lowest obtained values are reported. Analytes were quantified using internal standard methodologies. The extract's PAH portion was eluted through a silica/alumina column with methylene chloride. The collected PAH fraction was exchanged into hexane and concentrated to 250 µL in the same manner as the CH fractions.

6.0 TOTAL ORGANIC CARBON (TOC) ANALYSIS OF SEDIMENTS

6.1 Summary Of Methods

This procedure uses an elemental analyzer to determine the amount of total organic carbon in sediments. Samples were placed in vials and treated with 1N HCL to decompose all carbonate. Treated samples were centrifuged for 15 minutes and supernatant decanted. Vials containing samples were filled with deionized water, vortexed, centrifuged, and pH checked until pH was between 6 and 7. Samples were dried at less than 55°C until completely dry (approximately 3 days). Dried sediments were homogenized in a ball mill, and weighed into aluminum sleeves (1-5 mg) to the nearest 1 µg. Sediments were analyzed for total organic carbon by use of a Control Equipment Corp. Model 440-XA Elemental Analyzer.

6.2 Sample Preparation

Samples were homogenized thoroughly by stirring with a clean stainless steel spatula. Approximately 10 ml of subsamples to be analyzed were placed in sterile 20 ml polyethylene scintillation vials. The subsample was as representative as possible. Spatulas used to stir and transfer sediment to vials were washed with deionized water and wiped dry with Kimwipes between samples. Approximately 8 ml 1N HCL were added and mixed with the sample. Samples were vortexed and centrifuged at 2850 rpm for 20 minutes. The pH of the sample was checked and the supernatant was decanted. The sample was washed repeatedly with deionized water and centrifuged until pH was between 6 and 7. Samples were dried in a drying oven at 55° C or less until completely dry (about three days). Two clean 1/4 inch stainless steel ball bearings were placed into vials containing the samples, and samples were homogenized in a ball mill for about 15 minutes until they were of even particle size. One to five mg of treated sediments were weighed into aluminum sleeves to the nearest 1 µg. Then,

sleeves were crimped with forceps and placed in nickel sleeves in the combustion wheel of the elemental analyzer.

6.3 TOC Analysis

TOC was determined through the standard operating procedure of the Model 240-XA elemental analyzer. Built-in software in the computer interfaced to the analyzer was used to compute carbon content of the samples.

7.0 GRAIN SIZE ANALYSIS OF SEDIMENTS

7.1 Summary Of Methods

These procedures used sieve, hydrometer, and microscopic techniques to determine particle size of sediment samples.

7.2 Sample Splitting And Preparation

Samples were thawed and thoroughly homogenized by stirring with a spatula. Spatulas were rinsed of all adhering sediment between samples. Size of the subsample for analysis was determined by the sand/silt ratio of the sample. During splitting, the sand/silt ratio was estimated and an appropriate sample weight was calculated. Subsamples to be analyzed were placed in clean, pre-weighed beakers. Debris was removed and any adhering sediment was washed into the beaker.

7.3 Wet Sieve Analysis (Separation Of Coarse And Fine Fraction)

Beakers were placed in a drying oven and sediments were dried at less than 55°C until completely dry (approximately three days). Beakers were removed from the drying oven and allowed to equilibrate to room temperature for a least a half-hour. Each beaker and its contents were weighed to the nearest 0.01 g. This weight minus the empty beaker weight was the total sample weight. Sediments in beakers were disaggregated using 100 ml of a dispersant solution in water (such as 50 g Calgon/L water) and the sample was stirred until completely mixed and all lumps disappeared. The amount and concentration of dispersant used was recorded on the data sheet for each sample. Sample beakers were placed in an ultrasonic cleaner for 15 minutes for disaggregation. Sediment dispersant slurry was poured into a 63 μm (ASTM #230, 4 phi) stainless steel or brass sieve in a large glass funnel suspended over a 1L hydrometer cylinder by a ring stand, and all fine sediments were washed through the sieve with water. Fine sediments were captured in a 1L hydrometer cylinder. Coarse sediments remaining in sieve were collected and returned to the original sample beaker for quantification.

7.4 Dry Sieve Analysis (Coarse Fraction)

The coarse fraction was placed into a preweighed beaker, dried at 55° to 65°C, allowed to acclimate, and then weighed to 0.01 g. This weight, minus the empty beaker weight, was the coarse fraction weight. The coarse fraction was poured into the top sieve of a stack of ASTM sieves having the following sizes: No. 10 (2.0 mm), 18 (1.0 mm), 45 (0.354 mm), 60 (0.25 mm), 80 (0.177 mm), 120 (0.125 mm), and 170 (0.088 mm). The stack was placed on a mechanical shaker and shaken at medium intensity for 15 minutes. After shaking, each sieve was inverted onto a large piece of paper and tapped 5 times to free stuck particles. The sieve fractions were added cumulatively to a pretared weighing dish, and the cumulative weight after each addition determined to 0.01g. The sample was returned to its original beaker, and saved until sample computations were completed and checked for errors.

7.5 Hydrometer Analysis (Fine Fraction)

Hydrometers used for the analysis were precalibrated using the techniques of Lewis (1984). A reference cylinder was filled with water and 100 ml of dispersant solution. Prior to the analysis, a hydrometer reading was taken for Cc, the composite correction for temperature, dispersing agent, and the meniscus.

For each of the sample cylinders, the volume was raised to 1000 ml using tap water. The hydrometer number was recorded, the temperature was noted, and the sample added and stirred for 1 minute.

Hydrometer readings were taken at 1 minute, 3 minutes, 10 minutes, 30 minutes, 90 minutes, 4.5 hours and 24 hours. If the water temperature had changed by greater than 2°C then hydrometer corrections were remeasured. The colloidal weight was determined by subtracting the other fractions from the total weight.

7.6 Analytical Procedures

Fractional weights and percentages for various particle size fractions were calculated. If only wet sieve analysis was used, weight of fine fraction was computed by subtracting coarse fraction from total sample weight, and percent fine composition was calculated using fine fraction and total sample weights. If dry sieve and hydrometer analysis was employed as well, fractional weights and percentages for sieve and hydrometer fractions were calculated using custom software on a Macintosh computer. Calibration factors for each hydrometer were pre-stored in the computer.

7.7 Microscopic Descriptive Analysis Of Particle Configurations

Two small (~ 1 ml) subsamples of sediments from undisturbed cores were prepared by saturation with alcohol and gentle disaggregation and examined under a dissecting scope at 25 - 40 magnification. Notable features were then assigned one of the following abundance categories (in increasing order of abundance): absent, very rare, rare, common, abundant, very abundant. Samples were sorted on the following categories: fecal pellets (~ 0.1, 0.2, 0.3, >0.4 mm), diatoms (chains and centric), plant material, worm tubes, shell fragments, foraminiferans. These observations were tabulated to assist in the interpretation of possible sediment grain size effects on infaunal organisms in the toxicity tests.

8.0 ANALYSIS OF BENTHIC COMMUNITIES

8.1 SUMMARY OF METHODS

Investigators from the San Francisco Estuary Institute (SFEI), the Department of Water Resources, and the City and County of San Francisco participated in a cooperative study in which sediment samples from three candidate reference sites were analyzed to determine benthic community structure (SFEI, 1997). Paradise Cove and the two San Pablo Bay sites were samples over a three year period using a 0.05 m^2 Ponar Grab. Sediments from the grab samples were sieved through a 0.5 mm screen to remove benthic macrofauna, which were preserved and identified to the lowest practical taxon. Classification analysis (Smith et al., 1988) was used to determine how species composition and abundance from candidate reference sites compared with those of 124 other samples collected from around the Bay/Delta. Analyses based on numbers of species and individuals at each station (rather than formal diversity indices) are described in greater detail by SFEI (1997). Additional analyses were conducted based on identifying species and higher taxonomic groups characteristic of impacted and non-impacted sediments. These anayses are described further in SFEI (1997).

9.0 QUALITY ASSURANCE/QUALITY CONTROL

9.1 Summary

Summaries of quality assurance and quality control procedures were described under separate cover in the Bay Protection and Toxic Cleanup Program Quality Assurance Project Plan (Stephenson et al., 1994). That document described procedures within the program which were in place to ensure data quality and integrity. In addition, individual laboratories prepared quality assurance evaluations of each discrete set of samples analyzed and authorized by task order. These documents were submitted to the California Department of Fish and Game for review, then forwarded to the State Water Resources Control Board. Data quality is described in the Results section of this report.

10.0 STATISTICAL ANALYSIS

10.1 Summary Of Methods

Analyses were performed to determine the statistical significance of relationships between sediment toxicity test results, contaminant concentrations, and various natural factors at reference sites. Descriptive statistics and graphics were used to present toxicity data from reference and test sites to assist in the evaluation of test performance and reference site selection. Toxicity data from reference sites were also used to calculate tolerance limits to be used as a relative standard against which to compare toxicity data from test sites.

10.2 Determining Significant Relationships

Spearman rank correlations were used to evaluate the statistical significance of associations between sediment toxicity test results, contaminant concentrations, and various natural factors at reference sites. Toxicity data were analyzed in relation to synoptic measurements of sediment grain size, TOC, trace metal and trace organic contaminant concentrations, and in relation to ammonia and hydrogen sulfide measured during the toxicity tests, as described in Sections 2.1 and 2.2.

10.3 Descriptive Statistics

10.3.1 *Chemical Data*

The degree of chemical contamination at each site was characterized by averaging ERM and PEL quotients. ERM (Effects Range Median) and PEL (Probable Effects Level) values have been derived for 32 chemicals or chemical classes by examining a large number of previous studies to determine associations between chemical concentrations and adverse biological effects (Table 8). The derivation and application of ERM and PEL values have been previously described (Long et al., 1995, 1998; McDonald, 1994). These studies have indicated that adverse biological effects are probable when chemical concentrations in test sediments are higher than the ERM or PEL values. Concentrations of these chemicals measured in samples from the present study were divided by their respective ERM or PEL values to derive ERM or PEL quotients for each chemical. ERM and PEL quotients for all available chemicals were then averaged to give a relative measure of overall pollution at each site.

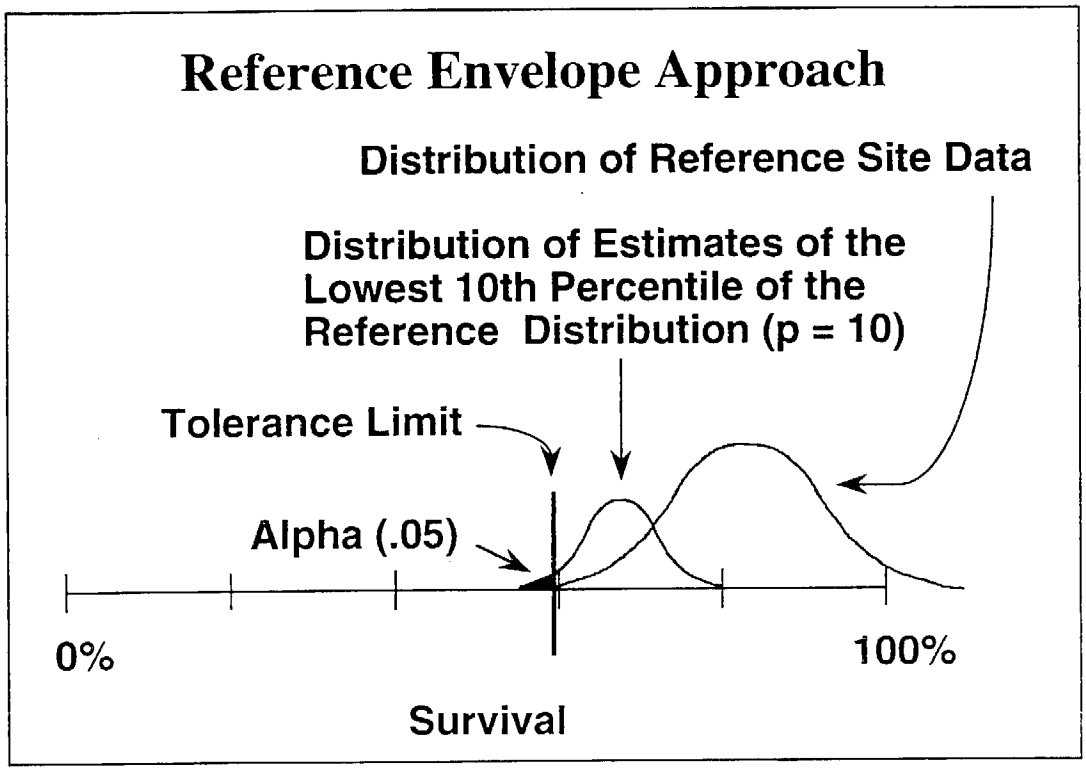
Table 8. Sediment Chemistry Guidelines
Developed by NOAA and the State of Florida

SUBSTANCE	State of Florida (1) TEL	State of Florida (1) PEL	NOAA (2) ERL	NOAA (2) ERM
Organics (ug/kg- dry weight)				
Total PCBs	21.550	188.79	22.70	180.0
PAHs				
Acenaphthene	6.710	88.90	16.00	500.0
Acenaphthylene	5.870	127.89	44.00	640.0
Anthracene	46.850	245.00	85.30	1100.0
Fluorene	21.170	144.35	19.00	540.0
2-methylnaphthalene	20.210	201.28	70.00	670.0
Naphthalene	34.570	390.64	160.00	2100.0
Phenanthrene	86.680	543.53	240.00	1500.0
Total LMW-PAHs	311.700	1442.00	552.00	3160.0
Benz(a)anthracene	74.830	692.53	261.00	1600.0
Benzo(a)pyrene	88.810	763.22	430.00	1600.0
Chrysene	107.710	845.98	384.00	2800.0
Dibenz(a,h)anthracene	6.220	134.61	63.40	260.0
Fluoranthene	112.820	1493.54	600.00	5100.0
Pyrene	152.660	1397.60	665.00	2600.0
Total HMW-PAHs	655.340	6676.14	1700.00	9600.0
Total PAHs	1684.060	16770.54	4022.00	44792.0
Pesticides				
p,p'-DDE	2.070	374.17	2.20	27.0
p,p'-DDT	1.190	4.77		
Total DDT	3.890	51.70	1.58	46.1
Lindane	0.320	0.99		
Chlordane	2.260	4.79	0.50	6.0
Dieldrin	0.715	4.30	0.02	8.0
Endrin			0.02	45.0
Metals (mg/kg- dry weight)				
Arsenic	7.240	41.60	8.20	70.0
Antimony			2.00	2.5
Cadmium	0.676	4.21	1.20	9.6
Chromium	52.300	160.40	81.00	370.0
Copper	18.700	108.20	34.00	270.0
Lead	30.240	112.18	46.70	218.0
Mercury	0.130	0.70	0.15	0.7
Nickel	15.900	42.80	20.90	51.6
Silver	0.733	1.77	1.00	3.7
Zinc	124.000	271.00	150.00	410.0

(1) D.D. MacDonald, 1994

(2) Long et al., 1995

Figure 6. Schematic illustration of the reference envelope method for calculating tolerance limits. The tolerance limit in this illustration is the point at which there is 95% certainty that lower values are as low or lower than the 10th percentile of the reference site distribution of toxicity test results.



10.3.2 *Toxicity Data*

Toxicity data were analyzed by station and by site. Station means were derived from laboratory replicates of each individual sample. There were three stations (= field replicates) sampled at each site. Site means were derived from the three station means at each site. Individual station and site mean values were calculated for each sampling event (Tables 1 and 2).

To allow equitable comparisons of toxicity data among sampling events, all mean toxicity values were normalized to the negative laboratory control values for each series of tests. Samples that were salinity-adjusted were normalized to brine controls. To normalize sample mean toxicity values, they were simply divided by the mean value from the corresponding laboratory control and presented as a percentage of the control. These normalized toxicity data were used in all subsequent analyses.

10.4 Toxicity Comparisons Using the Reference Envelope Approach

10.4.1 *The Basic Tolerance Limit Concept for Toxicity Data*

One of the primary objectives of the Bay Protection and Toxic Cleanup Program (BPTCP) is the identification of specific areas of water and sediment quality concern, where adverse biological impacts are observed in areas with locally elevated concentrations of pollutants. Identification of problem sites is an essential step in prioritizing efforts to improve sediment and water quality through regulation and cleanup programs. The BPTCP efforts are focused on localized areas that are significantly more toxic than the larger surrounding area of the water body. In this study, we have employed a "reference envelope" statistical approach (Smith, 1995) to make such a distinction in San Francisco Bay.

The reference envelope approach uses data from "reference sites" to characterize the response expected from sites in the absence of severe localized pollution. Using data from the reference site population, a tolerance limit is calculated for comparison with data from test sites. Samples with toxicity values greater than the tolerance limit are considered toxic relative to the optimal ambient condition of the Bay.

The tolerance limits were calculated using station data (from individual field replicates at a site) rather than site data (means of field replicate stations from within a site), because it was anticipated that the tolerance limits would be used for comparison with individual samples (rather than field replicate means) from test sites.

Tolerance limits were calculated using reference site data collected during this study, and additional tolerance limits were calculated using an expanded data set. This expanded data set included data from this study, plus data from the same reference sites sampled during BPTCP screening surveys of San Francisco Bay, plus data from additional sites sampled for the SF Bay Regional Monitoring Program (RMP) that could potentially be used as reference sites. The BPTCP screening studies produced 11 additional data points for *Eohaustorius* tests in homogenized sediment, eight additional data points for sea urchin larvae tested in pore water, and four additional data points for sea urchin larvae tested at the sediment-water interface. The additional RMP sites were Pinole Point (RMP site BD30) in San Pablo Bay, Horseshoe Bay (RMP site BC21) in Central San Francisco Bay, and San Bruno Shoal (RMP site BB15) in South San Francisco Bay. The location and description of these sites is given by SFEI (1997). These RMP sites produced three, seven and six additional data points, respectively, for *Eohaustorius* tests in homogenized sediment

This relative standard established using reference sites is conceptually different from what might be termed the absolute standard of test organism response in laboratory controls. Rather than comparing sample data to control data using t-tests, with laboratory replication used to characterize the variance component (e.g., Schimmel et al., 1994), the reference envelope approach compares sample data against a percentile of the reference population of data values, using variation among reference sites as the variance component. The reference envelope variance component, therefore, includes variation among laboratory replicates, among field replicates, among sites, and among sampling events.

The reference stations are assumed to be a random sample from an underlying population of reference locations that serve as a standard for what we consider non-impacted conditions. The toxicity measured at different reference locations will vary due to the different local conditions that can affect the toxicity results. In order to determine whether sediments from a test location are toxic, we compare the bioassay results for the test location with the bioassay results from the population of reference locations.

If we assume that the toxicity results from the population of reference locations are normally distributed, then we can estimate the probability that the test sediment is from the underlying reference station distribution. For example, if the result for a test sediment was at the tenth percentile of the underlying reference location distribution (in the direction of toxicity), then we would know that there was about a 10% chance that the test sediment was from the distribution of reference locations.

However, we do not know the exact toxicity level at the tenth percentile of the reference distribution because we only have limited samples from the underlying distribution. We can only estimate where the tenth percentile lies. If we were to estimate the value of the tenth percentile a large number of times using different random samples from the reference distribution, we would obtain a (non-central t) distribution of estimates, with the distribution mode at the actual tenth percentile (Figure 6). In Figure 6, it can be seen from the distribution of estimates that about one half of the time the estimate from the sample will be above the actual tenth percentile. Ideally, we would like to identify an estimated toxicity value that would cover the actual tenth percentile for a large percentage of the estimates (say 95% of the time). We can obtain such a value from the left tail of the distribution of estimates where 5% of the estimates are less than the chosen value. We define "p" as the percentile of interest, and alpha as the acceptable error probability associated with an estimate of the pth percentile. Thus, in our example, p=10 and alpha = 0.05.

10.4.1.1 *Calculation of a Tolerance Limit using Naive Variance*

The following tolerance limit calculation is valid for studies in which there is a single source of variance, and the calculation utilizes a variance term refered to as the "naive variance." We can compute the toxicity level that will cover the pth percentile 1 minus alpha proportion of the time as the lower bound (L) of a tolerance interval (Vardeman, 1992) as follows:

$$L = X_r - [\, g_{\alpha,p,n} * S_r \,] \qquad (1)$$

where X_r is the mean toxicity result from the sample of reference stations, S_r is the standard deviation of the toxicity results among the reference stations, and n is the number of reference stations. The g values can be obtained from tables in Hahn and Meeker (1991) or Gilbert (1987). "S" contains the within- and between-location variability expected among reference locations. If the reference stations are sampled at different times, then S will also incorporate between-time variability.

We call L the "edge of the reference envelope" because it represents a cutoff toxicity level we will use to distinguish toxic from non-toxic sediments (Figure 6). The value used for p, and the resulting tolerance limit L, will depend on the level of certainty needed for a particular regulatory situation. In this study we choose multiple p values for evaluation of the method.

10.4.2 *Computation of Parametric Tolerance Limits with Multiple Sources of Variance*

The tolerance limit calculations described in Section 10.4.1.1 above are valid for studies in which there is a single source of variance. For the present study, and for most sediment monitoring study designs, there are four pertinent sources of random variance affecting a single measurement: variance due to time (sampling event), space (station), time by space interaction, and error (within time-space replicate variance). In terms of an ANOVA model, time and station are considered crossed main factors, and are treated as random factors since we wish to generalize the results to the larger population of all possible sampling events and stations in reference locations of the Bay. Presently, there are no methods available in the statistical literature for computing tolerance limits with such a model. This is probably due to the fact the distributional theory on the variance components for a crossed random model is lacking (Searle et al., 1992). Davis (1994) discusses using tolerance limits with a similar statistical model, but provides no guidance on how the method can be applied to actual data.

In such situations where computational formulae are not available for an inferential approach, bootstrap simulation methodology can be applied (Efron and Tibshirani, 1993). To compute the tolerance limits for the present study, we applied a parametric bootstrap method that simulates the sampling process, starting with population mean and variance components estimated from the study data. Using bootstrapping techniques, we generated values for $K_{p,\alpha}$, which is the bootstrap analog to the g statistic in formula (1) in the previous section. This value is then inserted into the previous formula to generate tolerance limits for applications where multiple sources of variance affect each measurement:

$$L = X_r - [\, K_{p,\alpha} * S_r \,] \qquad (2)$$

where, as before, X_r is the mean toxicity result from the sample of reference stations, S_r is the standard deviation of the toxicity results among the reference stations, and $K_{p,\alpha}$ is obtained by using bootstrapping techniques described below.

10.4.2.1 *Bootstrapping Procedures For Deriving $K_{P,\alpha}$ Values*

The method described below for computing $K_{P,\alpha}$ requires the estimations of the population variance components and the mean. The method used for variance component estimation was the Henderson Method I estimators for a 2-way crossed random model (Searle et al., 1992, p. 434). This model is appropriate for unbalanced designs and can be rapidiy computed (as is required for the proposed intense simulation approach). For the present application, variance components were estimated for survey and station (main factors), survey by station interaction, and error. The population mean was estimated as the arithmetic average of all the data values. In all calculations, surveys within two months of each other were considered the same survey. This procedure reduced the number of empty cells in the analysis.

A parametric bootstrap method (Efron and Tibshirani, 1993) was used to compute the $K_{P,\alpha}$ values in formula (2). The algorithms used to compute $K_{P,\alpha}$ are described in Figures 7 and 8. The symbols used in the Figures 7 and 8 are defined in the accompanying legend.

Given initial estimates of the population mean and variance components, Algorithm A (Figure 7) could be used to estimate a value for K_{P,α_a}. This algorithm is similar in concept to the algorithm used by Davies and Gather (1993) to compute a constant (such as $K_{P,\alpha}$) for a robust outlier detection technique that is in principle very similar to a tolerance interval. In general, algorithm A first computes a target P^{th} quantile value from the initial mean and variance component estimates. Note that the population standard deviation (SD) is the square root of sum of the individual variance components (Davis, 1994). Next, again using the initial means and variance component estimates, multiple sets of simulated data are produced (details of the data simulation process are given in the next section). For each set of simulated data, a population mean and standard deviation are estimated ($\bar{\mu}_s$ and $\bar{\sigma}_s$). Finally, given the multiple sets of simulated means and standard deviations, a K value is found such that the resulting bounds cover the target quantile value for a 1-α_a proportion of the simulations.

If the original survey data are used to estimate the initial means and variance components for algorithm A, simulations show that the resulting K will tend to produce coverage of the P^{th} quantile

of the parent population at a rate *lower* than the desired 1-α (or the rate of non-coverage will be greater than α). Bootstrap *calibration* (Efron and Tibshirani, 1993) can be used to provide an adjustment for producing a coverage closer to the desired level. Here bootstrap calibration involves bootstrapping bootstrapped results to estimate the *actual* coverage associated with a particular α_a value. If the coverages for a series of α_a values is computed, then the K value for the α_a where the actual coverage is approximately equal to *1*-α can be used instead of the K value for α.

Legend. Definitions of symbols used in Figures 7 and 8.

Symbol	Definition
P	P value used for $K_{P,\alpha}$ in all algorithms (P> 0.5 in simulations - note: $K_{P,\alpha} = K_{1-P,\alpha}$)
α_a, α_b	α value used for $K_{P,\alpha}$ in algorithms a and b, respectively
f	Number of random factors in statistical model
s	Simulation counter used within each algorithm
S_a , S_b	Number of simulations in simulation loop for algorithms a and b, respectively
$\hat{\sigma}^2_{ai}, \hat{\sigma}^2_{bi}$	Initial estimate of variance component for random factor i in algorithms a and b, respectively *(i=1 to f)*
$\hat{\sigma}^2_{si}$	Estimate of variance component for random factor i for simulation s *(i=1 to f)*
σ^2_i	Variance component for random factor i *(i=1 to f)*
$\hat{\sigma}_s$	Estimate of population standard deviation for simulation s
$\hat{\mu}_a, \hat{\mu}_b$	Initial estimate of population mean for algorithms a and b. respectively
$\hat{\mu}_s$	Estimate of population mean for simulation s
z_P	P^{th} quantile of the N(0,1) standard normal distribution
$\hat{Q}_a, \hat{Q}_b$	Estimate of the P^{th} quantile given the initial mean and variance estimates in algorithms a and b, respectively

Figure 7. Flow chart for the initial calculation of $K_{P,\alpha}$.

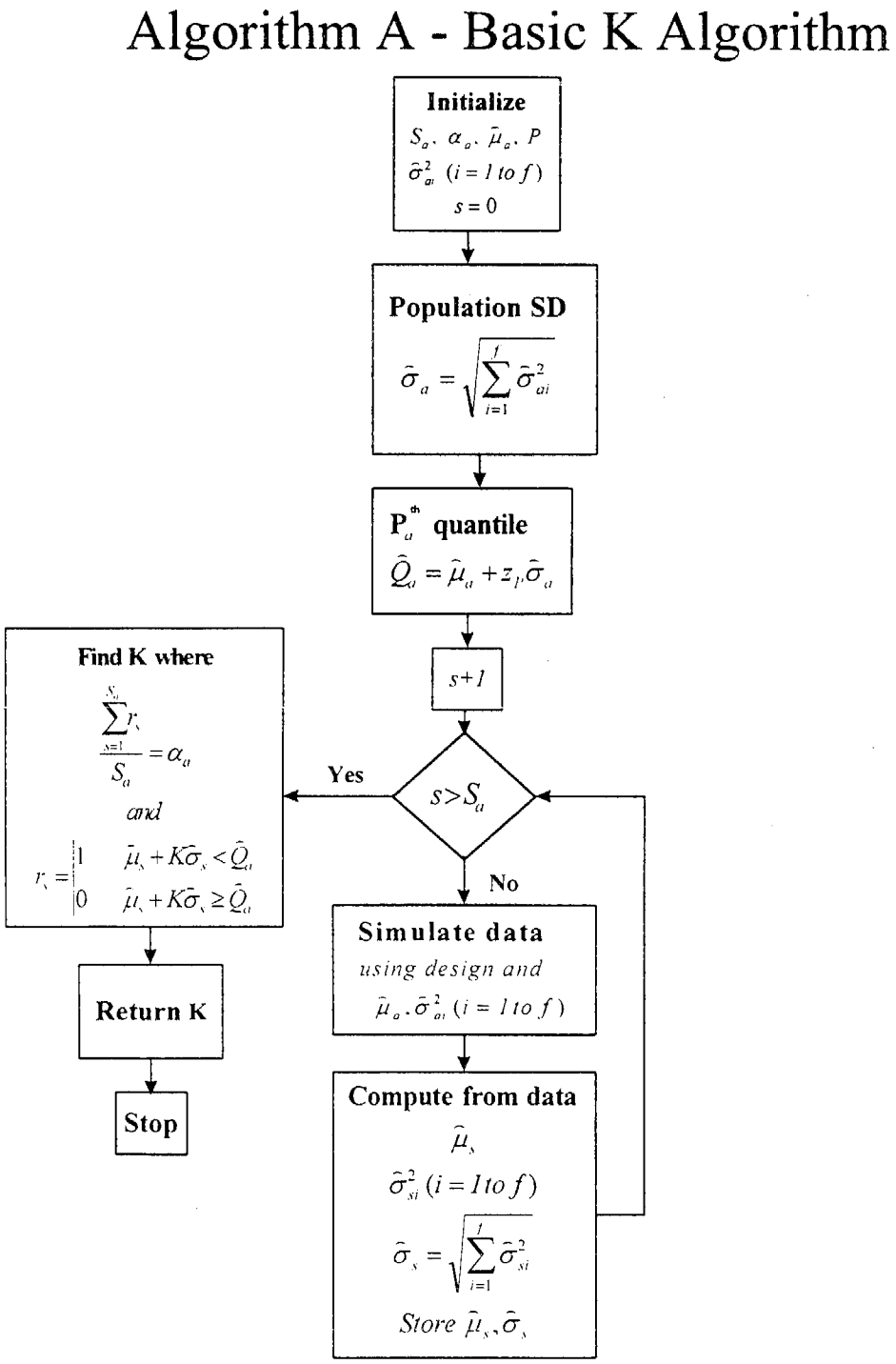
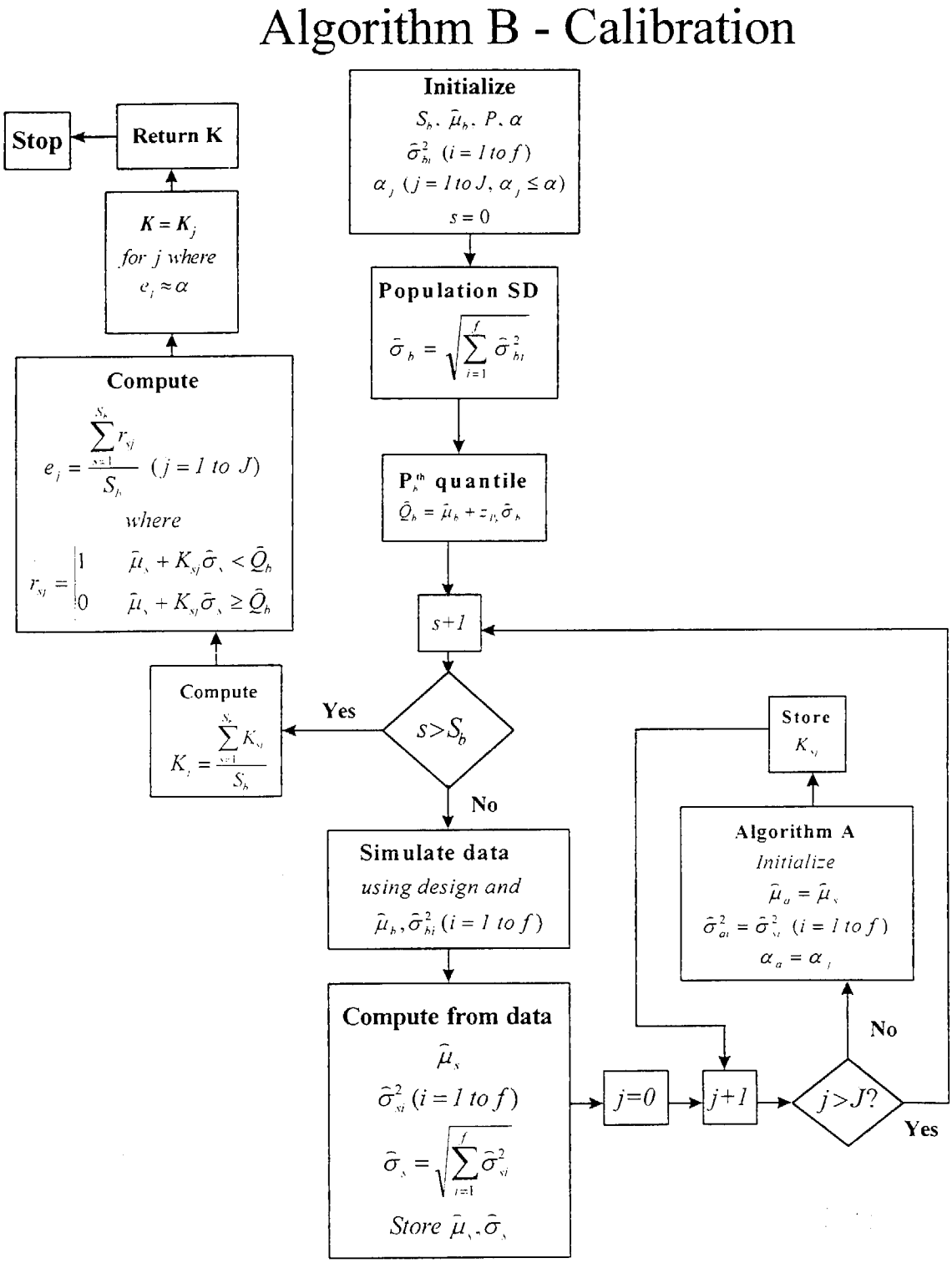


Figure 8. Flow chart for $K_{P,\alpha}$ calibration.



The calibration process (algorithm B) is described in Figure 8. First a target P^{th} quantile value ($\hat{Q}_b$) is computed from initial estimates of population mean and variance components. Then multiple sets of data are simulated from the initial mean and variance components. For each set of simulated data, K values for a series of α values (α_j ($j = 1$ *to* J)) are computed using algorithm A. After completion of the simulation loop ($s=1$ *to* S_b), the K_j value associated with each α_j is computed as the mean of the K_{sj} values over all simulations. Finally, the coverage of the target P^{th} quantile is computed for each α_j. The K_j associated with the (non)coverage (e_j) closest to α is the K value to use in formula (2). Note that the series of α_j values are only less than or equal to α. This is to prevent any calibration that might lead to even lower coverage than that associated with α.

10.4.2.2 *Data Simulation*

Both algorithms involve generating data from an overall mean and a set of four variance components. The respective variance components are for survey ($\hat{\sigma}^2_{time}$), station ($\hat{\sigma}^2_{space}$), survey by station interaction ($\hat{\sigma}^2_{time \times space}$), and error ($\hat{\sigma}^2_{error}$) This section briefly describes how the data are generated from the mean and variance components.

A multivariate random normal generator (Johnson, 1987) is used to produce cell means in the sampling design. For a simulation, let M be a n_s x n_t matrix of cell means for the crossed design, where n_s = # stations and n_t = # surveys. If m_i is the i^{th} column of M, then

$$m_i = AY_i + X,$$

where Y_i is a column vector of N(0,1) standard random normal deviates generated separately for column i. X is a column vector of $N(\mu, \hat{\sigma}^2_{space})$ random normal deviates, and A is a n_s x n_s matrix such that $AA' = \Sigma$. Here Σ is a n_s x n_s variance-covariance matrix with $\hat{\sigma}^2_{time} + \hat{\sigma}^2_{time \times space}$ in the diagonal and $\hat{\sigma}^2_{time}$ in the off-diagonal. Matrix A is computed by Choleski factorization of Σ. Once the cell means in M are computed, the replicate values in each cell are

simulated. A data value is simulated as a $N(m_{ij}, \bar{\sigma}^2_{error})$ random normal deviate, where m_{ij} is the cell mean i^{th} row and the j^{th} column of M. The number of replicates simulated in each cell equals the number of replicates in the sampling design.

10.4.3 *Computation of Nonparametric Tolerance Interval Bounds*

If the reference data are far from normal and cannot be transformed to approximate normality, then a *nonparametric* tolerance interval is more appropriate. To compute nonparametric tolerance intervals, we used the method proposed by Woodward and Frawley (1980), which is based on a method originally proposed by Hanson and Koopmans (1964). Woodward and Frawley (1980) show with simulations that their method works well with distributions that are skewed to the right (as is often the case with sediment chemistry data, though not with the currently evaluated sediment toxicity data). Here, a lower tolerance limit (L) is computed as

$$L = Y_n - b(Y_n - Y_1), \qquad (2)$$

where n is the number of sampling units, Y_1 is the smallest data value, Y_n is the highest data value, and b is a value dependent on n, P, and α that can be found in Table 2 in Woodward and Frawley (1980). This method has the advantage of working well with smaller sample sizes, in contrast with the more standard method based on the binomial distribution (Hahn and Meeker 1991, Chapter 5), which requires large sample sizes in most cases. In the results, nonparametric tolerance limits are shown for $P=.90$, $.925$, $.95$, $.97$ and $.99$ with $\alpha = .05$. Limits for $P=.925$ were approximated as averages of the limits for $P=.90$ and $P=.95$, and limits for $P=.97$ were approximated as averages of the limits for $P=.95$ and $P=.99$. The limits for $P=.925$ and $P=.97$ were estimated in this manner because values for b with $P=.925$ and $P=.97$ were not included in Table 2 in Woodward and Frawley (1980).

The computation of nonparametric tolerance intervals with the present crossed random statistical model is similar in concept to computing parametric tolerance intervals using the naïve variance

as the population variance. For the present application, the naïve variance would be computed using the standard variance formula without regard to the station, survey, interaction, and error variance components (see section 10.3.2.1 for how the variance components are used to compute the population variance or standard deviation). Davis (1994) shows that tolerance limits based on the naïve variance will tend be too liberal (i.e., the interval bound will cover the P^{th} quantile less than $1-\alpha$ proportion of the time). One way to counteract this tendency is to use a higher P for nonparametric tolerance limits than for parametric tolerance limits.

10.4.4 *Removal of Outliers*

The tolerance interval is a tool for screening toxicity results for values unlike that found in reference locations. Occasionally, unusually large effects are observed with no obvious explanation. If such results are included in the tolerance interval computations, the tolerance interval bounds can be so low that the method no longer has any utility as an environmentally-protective screening tool. To avoid this situation, outliers were removed from the data, as described below.

Eight data points, out of a total of 238 data points used in the analysis, were determined to be outliers and were dropped from the analysis. Of these, three outliers were indicative of extremely low toxicity, and five were indicative of extremely high toxicity. Three of the outliers came from experimental protocols with amphipods in pore water or intact cores. One (of 59) came from amphipod *Eohaustorius* solid-phase homogenate tests, and four (of 37) came from sea urchin larval tests in pore water. One sea urchin outlier had very low toxicity, three had high toxicity.

Box plots (Tukey, 1977) were displayed to identify the outliers for each bioassay test. In a box plot the distribution of data values is summarized; two features of the plot are relevant here. First, a central box with bottom and top edges at the 25th and 75th percentiles of the data distribution is displayed. Second, extreme outliers are identified as values found more than three interquartile ranges from the edge of the central box. Extreme outliers identified in this manner were removed from the data before computation of the tolerance intervals.

Results

1.0 DATA QUALITY

1.1 Chemistry Data

All trace metal and trace organic chemistry data presented in this report met or exceeded quality assurance guidelines, as outlined in the Bay Protection and Toxic Cleanup Program (BPTCP) Quality Assurance Project Plan (Stephenson, et al., 1994). QA data reports were prepared specifically for the data presented in this report, and were submitted to the SWRCB.

1.2 Grain Size Data

All sediment grain size data presented in this report met or exceeded BPTCP quality assurance guidelines, as outlined in the BPTCP QAPP. A QA data report covering grain size analysis was submitted to the SWRCB.

1.3 Toxicity Data

Not all of the toxicity data presented in this report met all QA criteria as outlined in the BPTCP QAPP. Deviations from QA criteria are described briefly here and in detail in the QA report for toxicity data reproduced in Appendix B. All deviations from toxicity test QA criteria were considered minor, and were not expected to affect interpretation of the data for the objectives of this study.

1.3.1 *Ampelisca Tests*

In two of the sampling periods, the amphipod survival in negative controls was less than the 90% criterion. In Spring of 1994, control survival was 85 ± 12%; in Fall of 1994, control survival was 81 ± 10%. Both of these tests were conducted with amphipods collected from San Francisco Bay. In the Spring, 1995 test, amphipods from Rhode Island were used, and control survival was 91 ± 9% and 96 ± 6% for the two sampling events during that season. The ability to meet the control survival criterion was one of the factors considered in this study's evaluation of toxicity tests for use in San Francisco Bay.

1.3.2 *Mytilus Porewater Tests*

In the test of samples collected in September, 1994, the percentage of normally developed larvae in negative controls was below the acceptability criterion of 70%. Data from those samples are presented only in the evaluation of test performance, not in the evaluation of reference sites or calculation of reference envelope tolerance limits. All *Mytilus* data presented for these purposes met the control acceptability criterion.

1.3.3 *Nebalia Tests*

The *Nebalia* test was undergoing preliminary development, and incorporated test acceptability criteria from the ASTM amphipod protocol (e.g., 90% control survival; ASTM, 1993). The test was conducted four times, and met the control response acceptability criterion only once (9/94). The 3/95 test had control survival of 85 ± 4%. While this was below the 90% criterion, this deviation was considered minor for the objectives of this study, and the data from that test were included in calculations for this report. Poor organism condition resulted in poor control response in the 4/94 and 4/95 tests, and data from these were not used, except in the evaluation of test performance (see Section 3.1, below).

1.3.4 *General*

Dissolved oxygen (DO) measurements indicated DO above 100% saturation in a number of tests (Appendix B). This deviation was within allowable measurement error (10%) in the majority of cases and was not expected to have affected test results. All other QA deviations involved minor salinity fluctuations or *Ampelisca* laboratory holding times of less than 48 hours, as described in Methods Section 3.3 and Appendix B. None of these deviations were expected to have significantly affected test results.

2.0 EVALUATION OF REFERENCE SITES

2.1 Measured Chemistry at Reference Sites

Chemical concentrations were compared to probable effects levels (PELs; MacDonald, 1994; Long et al., 1998) and effects range median values (ERMs; Long et al., 1995; 1998). PEL and ERM values are informal (nonregulatory) benchmarks to aid in the interpretation of sediment chemistry data (Long et al., 1998; Table 8). They were derived as mid-range points within the distributions of chemical concentrations associated with measures of adverse effects (ERMs) or associated with both effects and no-effects data (PELs). Only those chemicals for which PEL and/or ERM values have been derived were used in this analysis (see Methods section 10.2.1). Chemical concentrations exceeding ERM and/or PEL values do not necessarily indicate that biological effects will be observed in a given sample, but these guidelines are useful for evaluating the reference site data relative to previous studies.

PEL values for chromium, and PEL and ERM values for nickel were exceeded at all reference sites (Table 9a). The mean value for trace metal PEL quotients for each site (excluding nickel) ranged from 0.28 to 0.37, while the mean value for trace metal ERM quotients for each site (excluding nickel) ranged from 0.16 to 0.27. The reasons for excluding nickel from

calculations of mean ERM and PEL values are considered in Discussion Section 1.0, as are the implications of ERM and PEL comparisons for chromium and DDT compounds.

The mean PEL and ERM quotients for organic chemicals, including the elevated DDT values, ranged from 0.09 to 0.11 and 0.05 to 0.07, respectively. Organic chemical concentrations in reference site samples were generally well below guideline values, with two exceptions. A sample from Paradise Cove collected in March, 1995, matched the PEL value for dibenz[a,h]anthracene, a polycyclic aromatic hydrocarbon (Table 9b). The San Pablo Bay Island # 1 sample from Spring, 1995, had measured concentrations of p'p'DDT and total DDT that exceeded PEL and/or ERM values. While an ERM value for p'p'DDT has not been derived, the measured p'p'DDT concentration was 12.2 times the PEL value. However, the high p,p'-DDT concentration and the high p,p'-DDT to p,p'-DDE ratio observed in the San Pablo Bay, Island # 1 sample appeared anamalous. Therefore, a replicate analysis of the sample was performed. This replicate analysis produced similar PCB and PAH profiles as the initial sample, but failed to reproduce the high p,p-DDT result (p,p-DDT was not detected in the replicate analysis). It appears that the pesticide residues in this sample were subject to a higher degree of variability, which may have been a result of either isolation of pure p,p'-DDT within small sediment particles, or of decomposition of these residues after the initial analysis. A small particle may have had DDT embedded inside it where it was not bioavailable and did not degrade into either DDE (aerobic) or DDD (anerobic). However, failure to reproduce the initial measurement indicates that the DDT was not widely distributed in the sample. Good PCB and PAH reproducibility indicates that these residues were more evenly distributed in the sample.

The mean PEL quotient for all chemicals for which guideline values exist (trace metal and organic), excluding nickel, ranged from 0.28 to 0.37, while the mean ERM quotient for all chemicals ranged from 0.16 to 0.27. The mean ERM quotients for the three Regional Monitoring Program (RMP) sites that were included in Reference Envelope tolerance limit calculations were: 0.092 (San Bruno Shoal), 0.108 (Horseshoe Bay), and 0.095 (Pinole Point).

2.2 Toxicity Test Results at Reference Sites

Samples from San Pablo Bay Island #1 generally showed little toxicity. Mean values for three field replicates were greater than 80% of control response in all tests except the September, 1994, test of intact sediment with *Eohaustorius*, and the March, 1995, *Neanthes* growth test (Figure 9). No individual field replicates from any other tests produced a value lower than

75% of control (Table 10; note that data in Table 10 are absolute values, and are not given as percentages of control values, as they are in all Figures).

Samples from San Pablo Bay Tubbs Island produced a similar pattern, but had lower levels of survival in solid phase tests with *Ampelisca*, *Eohaustorius*, and *Neanthes* (Figure 10). The same *Neanthes* growth test and intact sample *Eohaustorius* test produced poor results. The performance of these tests is discussed in the next section.

Paradise Cove samples showed little toxicity in homogenized sediment tests with amphipods, mussel and sea urchin larval tests in porewater, and sea urchin larval tests at the sediment-water interface (Figure 11). Results were more variable for intact sample and porewater tests with amphipods, and in tests with *Neanthes* and *Nebalia*, as will be discussed in the next section.

Patterns of response in the various toxicity tests at the North and South sites from South San Francisco Bay were similar (Figures 12 and 13). With the exception of amphipod porewater results that reflect relatively low survival in test controls, results were consistently between 80% and 100% of the control response.

Notable in the data from Tomales Bay, Marconi Cove, is the relatively poor survival of the amphipods *Eohaustorius* in homogenized sediment (Figure 14). Results from the other tests were comparable to those obtained from the other reference sites. As will be discussed in following sections, the Tomales Bay site had the highest percentage of clay particles, with greater than 60% of the sample mass composed of particles less than 4 μm.

With the exception of *Eohaustorius* tests with intact samples, toxicity test results from Bolinas Lagoon were consistently greater than 80% of control response (Figure 15).

2.3 Variability Among Field Replicates at Reference Sites

With the exception of porewater and intact core tests with the amphipods, variability among reference site field replicates was relatively low (as indicated by error bars in Figures 9 through 15; see also Table 10). The highest coefficient of variation (CV = standard deviation divided by the mean) for field replicate variability was 33% for *Eohaustorius* in homogenized sediment from Tomales Bay, while other high values included 26% for *Neanthes* survival at Tomales Bay and 25% for *Eohaustorius* at the South South Bay site.

2.4 Temporal Variability at Reference Sites

Temporal variability is indicated by differences between adjoining histogram bars in Figures 9 through 15. Again with the exception of porewater and intact core tests with amphipods, temporal variability in toxicity test results at reference sites was relatively low. The highest temporal variability was among results of *Neanthes* growth tests at Tubbs Island and Island #1 in San Pablo Bay (CVs = 35% and 24%, respectively), and among results of *Neanthes* survival at Tomales Bay (CV = 27%).

2.5 Physical Characteristics at Reference Sites

2.5.1 *Grain Size*

Sediments from candidate reference sites were generally fine-grained (Figures 16 through 22). Samples from nearly all sites had a broad distribution of particle sizes ranging from approximately 0.2 μm to 60μm (colloids/clays to silts). The most abundant size fractions were generally in the 1 to 4 μm range (clay) at nearly all sites, and there was moderate temporal variation in grain size at all sites. Bolinas Lagoon samples tended to have a broader range of particle sizes, with a greater fraction of silt and sand (Figure 22). Tomales Bay had the narrowest distribution, with clays and colloids often accounting for greater than 60% of the sample mass (Figure 21). Microscopic analysis revealed that Tomales Bay samples had a greater abundance of small fecal pellets, but did not differ from other sites in their abundance of diatoms, foraminiferans, plant material, worm tubes or shell fragments. Toxicity to *Eohaustorius* (in homogenized sediment) correlated significantly with the presence of clay/colloid particles at Tomales Bay (Spearman Rank Correlation, $n = 7$, $p < 0.05$). *Eohaustorius* toxicity in homogenized sediment also correlated significantly with the clay/colloid fraction at all sites, and *Neanthes* toxicity correlated significantly with the percentage of fine grained sediment (silt plus clay) at all sites (Table 11).

Grain size distributions at the reference sites were similar to those found at sites being investigated as candidate toxic hot spots, including Castro Cove and Islais Creek (Figure 23). Sediment sampled from Clipper Cove had a slightly bimodal distribution, with a moderate amount of medium-grained sand and a greater fraction of clay.

2.5.2 *Total Organic Carbon (TOC)*

Total organic carbon at the reference sites ranged from 0.74% to 2.39% (Appendix A, Section II). This was lower than that observed at Islais Creek (4.32%) and possibly at other sites where sludge or other sewage derived organic matter accumulate. Castro Cove (1.43%) and Clipper Cove (1.10%) were within the range of TOC values obtained at reference sites.

Tomales Bay had consistently higher TOC than did the reference sites within San Francisco Bay, averaging 2.29% compared to an overall reference site mean of 1.40%. TOC was significantly negatively correlated with survival of *Eohaustorius* in porewater, survival of *Eohaustorius* in homogenized sediment, and normal development of sea urchin larvae in porewater at reference sites (Table 11).

2.6 Benthic Community Analyses at Reference Sites

Assessments of sediment quality commonly include an analysis of benthic community ecology. Through cooperative efforts with the SF Bay RMP, three reference sites from this study were included in RMP pilot studies evaluating pollution impacts on benthic communities (SFEI, 1997). These efforts to classify sites are based on the presence or absence of species that are indicative of unimpacted sites, species indicative of impacted sites, taxonomic groups indicative of unimpacted sites (such as amphipods and echinoderms) and taxonomic groups indicative of impacted sites (such as oligochaetes and chironomids). During three years of sampling (1994 to 1996), Island #1, Tubbs Island, and Paradise Cove had 36%, 22%, and 10% impacted species, 9%, 11%, and 19% unimpacted species, and 3%, 11%, and 42% amphipods, respectively. There were insignificant numbers of echinoderms, oligochaetes or chironomids at all three sites. This preliminary data suggest that the benthic community of the Island #1 site may be moderately impacted by pollutants. Tubbs Island appears to have a less impacted fauna, and the assemblage observed at Paradise Cove appears to be indicative of an unimpacted benthic community. No data were available for the South Bay reference sites. A more extensive discussion of these results is presented in the RMP 1996 Report (SFEI, 1997).

The three RMP sites that were included in Reference Envelope tolerance limit calculations were also sampled for benthic community analyses (SFEI, 1997). During three years of sampling (1994 to 1996), Pinole Point, Horseshoe Bay, and San Bruno Shoal had 22%, 12% and 17% impacted species, 7%, 11%, and 15% unimpacted species, and < 1%, 27%, and 23% amphipods, respectively. This preliminary data suggest that the benthic community of Point Pinole may be moderately impacted by pollutants. The benthic communities of Horseshoe Bay and San Bruno Shoal do not appear to be impacted. These interpretations are preliminary, and are discussed further in the RMP 1996 Report (SFEI, 1997).

3.0 EVALUATION OF TOXICITY TESTS

3.1 Test Performance

3.1.1 *Acceptability of Test Control Response*

The degree to which each toxicity test met control acceptability criteria is indicated in Figures 24a and 24b. Control responses for each test, along with station means of all laboratory replicate toxicity data are given in Table 10. Solid-phase sediment tests with the amphipod *Eohaustorius* met control acceptability requirements in four of four trials (Figure 24a). Porewater and sediment-water interface (SWI) tests with the sea urchin *Strongylocentrotus* met control acceptability requirements in all trials (Figure 24b). All other tests had at least one trial in which control response was below the criterion. The *Ampelisca* test, as described above, met the criterion in both trials with east coast amphipods, but fell short of the criterion with amphipods collected in San Francisco Bay.

Tests with the highest percentage of test failures based on control acceptability were the *Nebalia* solid-phase test and the porewater test using *Eohaustorius*. Poor condition of cultured *Nebalia* test organisms (in one trial) and field collected organisms (in another trial) appeared to be responsible for those poor test results. There is no specific test acceptability criterion for porewater tests with *Eohaustorius*, because this infaunal amphipod is not routinely tested for 10 days in water only exposures. In four trials testing *Eohaustorius* in porewater, control survival was 80, 84, 48, and 84%, all below the 90% criterion established for amphipods tested in solid-phase sediment.

3.1.2 *Variability among Laboratory Replicates*

Variability among laboratory replicates of test samples is often used to define the variance component in statistical tests, and is therefore a primary factor affecting test power to discriminate among samples. It is used here as a measure of the consistency of response among test organisms. Tests using developing larvae of the sea urchin *Strongylocentrotus* had the lowest variability among laboratory replicates (Figure 25). Variability among intact sediment cores tested with *Eohaustorius* had the highest variability. *Neanthes* growth and survival and *Eohaustorius* survival in porewater also had higher than average variability among laboratory replicates.

3.1.3 *Test Sensitivity*

The ability to discriminate between sites with presumed low and high concentrations of measured chemicals was the primary indicator of test sensitivity in this study (Figure 26). Islais Creek and Castro Cove were used as examples of sites with high levels of pollution, though this is based on previous studies (Long et al., 1988; Flegal et al., 1994), as chemistry

was not measured at these sites in this study. Paradise Cove was used as an example of a reference site in this comparison because it is located between Islais Creek and Castro Cove, and because test responses were generally similar to those from the other four candidate reference sites within the Bay (Figures 9 through 13, Table 10).

Comparisons of toxicity data from Islais Creek, Castro Cove and Paradise Cove indicate that four tests demonstrated reduced survival or abnormal development at Islais Creek and/or Castro Cove, while two tests showed no difference between the sites. Islais Creek and Castro Cove samples produced significantly lower survival than controls in solid phase tests using the amphipods *Eohaustorius* and *Ampelisca*. Porewater and SWI tests using larval sea urchins exhibited significant toxicity at Islais Creek, but not at Castro Cove or Paradise Cove. Solid phase tests using *Neanthes* growth and survival and *Nebalia* survival produced high growth and survival at both reference and contaminated sites (Figure 26). No data were available for the *Mytilus* test at Islais Creek or Castro Cove due to less than acceptable control response (see Results Section 1.2.2).

Samples from Islais Creek had concentrations of total sulfide and unionized ammonia that may have been at least partially responsible for effects observed in some of the toxicity tests (Table 12). In tests with *Eohaustorius* and *Ampelisca*., presumed threshold levels of total sulfide were exceeded in test container sediment interstitial water, but not in overlying water. However, the mobile amphipods are capable of avoiding interstitial sulfide by emerging from test sediments or by inhabiting more highly oxidized surficial layers. For this reason, sulfide application limits have not been established for these tests (EPA 1994). The calculated total sulfide LOEC (lowest observed effect level) for development of sea urchin embryos was exceeded in Islais Creek interstitial water tested with this protocol (Table 12). Total sulfide toxicity thresholds, rather than those for the toxic hydrogen sulfide form, were used because literature comparative data were presented as total sulfide, and algorithms for calculating the percentage of hydrogen sulfide in seawater varied between laboratories. Hydrogen sulfide data, calculated according to methods described in Phillips et al (1997), are given in Appendix B, Section 1.

Unionized ammonia threshold values were exceeded in Islais Creek samples in tests with *Ampelisca*, sea urchins in porewater and sea urchins at the sediment-water interface (Table 12). While these threshold exceedences suggest that ammonia and sulfide may have been responsible for toxicity at Islais Creek, preliminary toxicity identification evaluations (TIEs) of concurrently collected Islais Creek samples indicate that substantial toxicity remained in the

samples after hydrogen sulfide and ammonia were removed (by aeration and zeolite treatment, respectively; Hansen and Assoc, 1996). Neither the Castro Cove nor the Paradise Cove samples had levels of hydrogen sulfide or ammonia above presumed threshold values for biological effects (Table 12).

3.1.4 *Relationship with Chemistry at Reference Sites*

Toxicity test response was significantly negatively correlated to concentrations of some measured chemicals at reference sites (Table 13). Survival of *Eohaustorius* in homogenized sediment was significantly negatively correlated to concentrations of arsenic and copper, while survival of *Eohaustorius* in porewater was significantly negatively correlated to concentrations of copper, iron, antimony, zinc, and p'p'DDE. Normal development of sea urchin larvae in porewater was significantly negatively correlated to concentrations of total PCBs, while normal development of sea urchin larvae at the sediment-water interface was significantly negatively correlated to concentrations of arsenic and p'p'DDE. The significance of these correlations is uncertain, however, because none of these chemicals exceeded ERM values (see Results Section 2.0), and there was minimal toxicity in samples from these sites, with the exception of survival of *Eohaustorius* in porewater.

3.1.5 *Relationship with Natural Factors*

Only survival of *Eohaustorius* in porewater was significantly negatively correlated with test solution ammonia or hydrogen sulfide at reference sites (Table 11). However, neither ammonia nor hydrogen sulfide concentrations were as high as those reported to be toxic to *Eohaustorius* (Appendix B and Table 12). Survival of *Eohaustorius* in porewater, survival of *Eohaustorius* in homogenized sediment, and normal development of sea urchin larvae in porewater were each significantly negatively correlated with total organic carbon (TOC). Survival of *Eohaustorius* and *Neanthes* in homogenized sediment were both significantly negatively correlated with grain size: *Eohaustorius* with percent clay/colloids (the finest measured fraction) and *Neanthes* with percent fines (the combined silt and clay fractions). As above, test organism survival and normal development were generally high at reference sites, except for survival of *Eohaustorius* in porewater (Table 10). These correlation analyses were part of the assessment of reference sites; data from presumed contaminated sites were not included. Ammonia, sulfide, grain size and TOC at suspected contaminated sites are discussed in Section 3.1.3.

3.1.6 *Overall Evaluation of Test Protocols*

Results from amphipod tests with intact sediment cores and sediment porewater were highly variable and subject to low control performance. These tests were intended for specific

applications other than routine monitoring. The intact core tests were conducted for the purpose of investigating the effects of homogenization on sample toxicity. Carnivorous annelids much larger than the test amphipods were occasionally observed in the intact core samples, and predation may have had a significant effect on test results. The porewater amphipod tests were conducted to provide screening data for Toxicity Identification Evaluations (TIEs). Lack of a sediment matrix is known to exert additional stress on test amphipods.

Nebalia tests were subject to poor control performance, and did not respond to sediments from test sites that were toxic to amphipods. This test was also experimental, conducted in an effort to develop a test with greater tolerance to hydrogen sulfide and ammonia. Neither of these compounds appeared to be a factor in *Eohaustorius* toxicity in the sediments tested, though ammonia may have been a factor affecting survival of *Ampelisca* in Islais Creek samples. *Neanthes* tests produced greater variability among field replicates and among laboratory replicates than did the amphipod tests (with the exception of porewater and intact core amphipod tests). *Neanthes* tests did not respond to sediments from test sites that were toxic to amphipods.

Tests with amphipods in homogenized sediment resulted in acceptable control performance, with the exception of *Ampelisca* collected in San Francisco Bay, as described above. Amphipod survival declined significantly in sediments from test sites, presumably responding to sediment pollutants. Past studies of Castro Cove have documented high levels of numerous pollutants, though no recent data was available to characterize the Islais Creek site.

With the exception of one set of tests with mussel larvae, larval development tests in porewater and at the sediment-water interface exceeded control acceptability criteria. Several factors often complicated the interpretation of larval porewater test results. Sulfide and/or ammonia were often measured at concentrations above toxicity thresholds in porewater samples (Table 12), making it difficult to determine the toxic effects of any available pollutants. Porewater salinity adjustment caused varying degrees of sample dilution, depending on the original salinity of the samples. This variable sample dilution made it difficult to compare test results between sites. Sea urchin larvae tested at the sediment-water interface (SWI) were generally exposed to lower concentrations of toxic sulfide and/or ammonia, with concentrations of these compounds often below threshold values when corresponding porewater sample concentrations were above thresholds (Table 12). SWI tests were not affected by original sample salinity, since all tests

used unadjusted overlying water from the same source, and data from the SWI tests were more directly comparable among sites.

4.0 TOLERANCE LIMITS BASED ON REFERENCE SITE TOXICITY DATA

4.1 Distribution of Reference Site Toxicity Data

The distributions of reference site toxicity data for all protocols are presented in Figures 27 through 39. Figures 29a, 34a, and 36a contain additional data from BPTCP screening surveys and from the SF Bay RMP (Table 14).

Three outliers were identified in the reference site toxicity data from this study. One of these outliers was from the test of intact sediments with amphipods, in which the outlier value was 20% of the test control value (Figure 30a). The other two were from tests of sediment porewater with amphipods, in which the outlier values were greatly in excess of the test control values, which were lower than acceptable in tests of solid-phase sediment (Figure 31a, Table 10). These tests, as described in Section 3.6 above, were experimental and subject to high variability. The expanded data set that included BPTCP screening data and RMP data had additional identified outliers. The outliers included one low value in the *Eohaustorius* test of homogenized sediment (Figure 29a), and three low values plus one high value in the sea urchin porewater tests (Figure 34a).

Reference site toxicity data from this study appear to be normally distributed (Figures 27 through 39), and there were no significant departures from normal distributions (alpha > 0.05). Combined data sets (including BPTCP screening and RMP data) were normally distributed after outliers were removed. Sea urchin larval tests had the lowest variability about the mean response (Figures 33 through 36), though outliers existed in the expanded data set (Figure 34a). Amphipod tests in homogenized sediment had intermediate distributions, in terms of variability within the data set (Figures 27a, 28a, and 29a), while *Neanthes* tests and tests of intact cores and porewater with amphipods had the greatest variability in response to reference site sediments.

4.2 Tolerance Limits for Sediment Toxicity

The amount of variability, the factors contributing to observed variability, and the mean response to reference site sediment exhibited by each protocol influenced the tolerance limits calculated for sediment toxicity in San Francisco Bay (Figures 27 through 39; Table 15). Tolerance limits are presented in a number of ways to demonstrate the effects of various factors affecting reference envelope tolerance limit calculations. Tolerance limits calculated using "naive variance"

(assuming a single source of variation), tolerance limits calculated using bootstrap simulations (to account for multiple sources of variation, such as exist in the present study with multiple sampling times and locations), and non-parametric tolerance limits are presented for a number of different data combinations and "p" values (see Methods Section 10.4).

Since the naive variance calculation assumes that all variance is random variance, tolerance limits calculated using this method will approximate those produced by the multiple-variance bootstrap simulation calculations when the error term is the primary variance component. This tends to be the case with data from the *Eohaustorius* tests of homogenized sediment (Figure 29b). The error variance component accounts for 58% of the total variance in the *Eohaustorius* homogenized sediment test data (Table 16), and the naive variance and bootstrap generated tolerance intervals are very similar. When the variance is spread more evenly among variance components, as is the case with *Eohaustorius* porewater and mussel larval tests (Figures 31b and 32b), the differences between tolerance limits calculated by the two methods is greatest. The non-parametric tolerance limit calculations are most influenced by the absolute spread in the data distribution, since this method depends on the range of values. Thus, for sea urchin porewater test values with outliers removed (Figure 34b) the non-parametric tolerance limits are similar to limits calculated with naive and bootstrap parametric methods, but when outliers are added and the range is extended (Figure 34c), the non-parametric tolerance limits are much lower.

The naive variance and non-parametric tolerance limits are presented for comparison, since the data are normally distributed (parametric) and are characterized by multiple sources of variation (so the bootstrap simulation method is appropriate).

Tolerance limits decrease with "p" value at various rates, depending on the total variation and distribution of variation among variance components. As p values decrease, tolerance limits proceed toward lower percentiles of the reference distribution. In cases where the reference site distribution has a high mean value and low overall variation about that mean, tolerance limits are relatively high (as in Figure 34b). When mean values are relatively high, but variability is high as well (as in Figures 37b and 38b), resulting tolerance limits may be low relative to previous interpretations of sediment toxicity (e.g., Swartz et al., 1985a; Schimmel et al., 1994). When mean reference site values are low and variability is high, tolerance limits are very low (as in Figure 30c). In such cases, negative tolerance limits are possible, and application of this method would deny any logical reason for testing, since any possible test result would surpass the limit.

Sample results from the two test sites (Castro Cove and Islais Creek) can be compared against calculated tolerance limits. The tolerance limit for the lowest 10th percentile of the reference site distribution (p = 10) for *Ampelisca* tests was 71% of the control (Table 15). Tests of Castro Cove and Islais Creek sediment produced survival rates of 36% and 67% of controls, respectively. Similar values for the *Eohaustorius* test were: 10th percentile tolerance limit 70%, Castro Cove 35%, and Islais Creek 60%. Solid-phase tests with *Neanthes* were above the tolerance limits for samples from both sites. Porewater tests of sea urchin larvae produced a 10th percentile tolerance limit of 94% of control response, compared to 104% at Castro Cove and 0% at Islais Creek (though sulfide and ammonia were at toxic levels in Islais Creek porewater; Table 12). Similar results were observed for sea urchin SWI tests, though sulfide and ammonia toxicity at Islais Creek are less probable.

Figure 9. Results of toxicity tests of samples from San Pablo Bay, Island #1. Each column represents a sampling event; error bars are ± one standard deviation among field replicates.

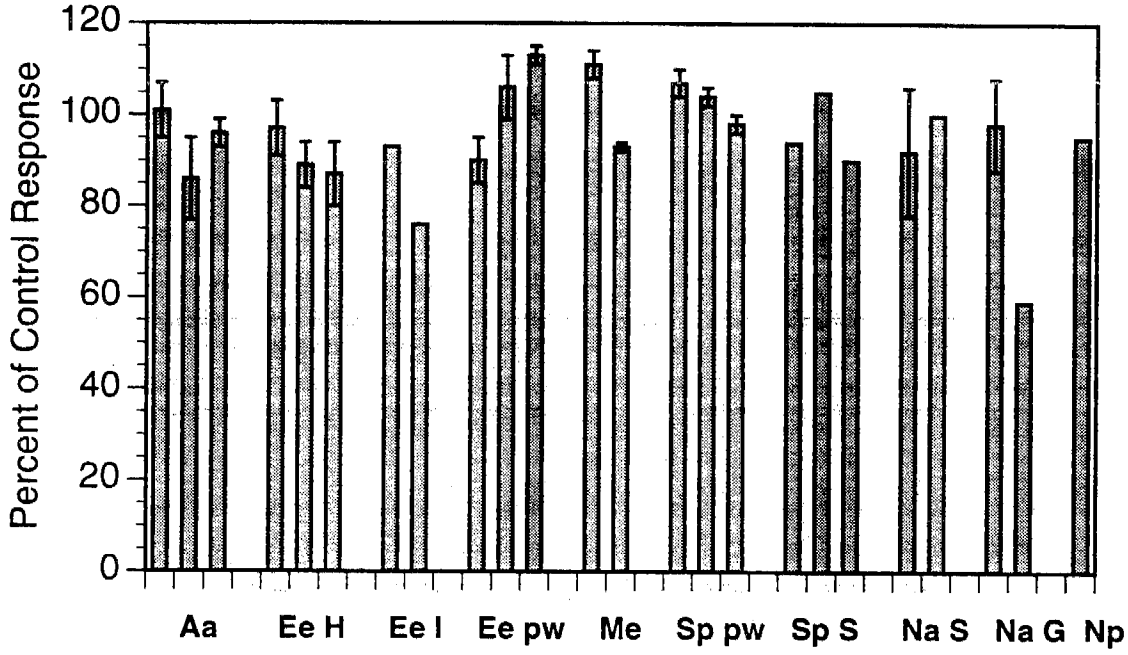


Figure 10. Results of toxicity tests of samples from San Pablo Bay, Tubbs Island. Each column represents a sampling event; error bars are ± one standard deviation among field replicates.

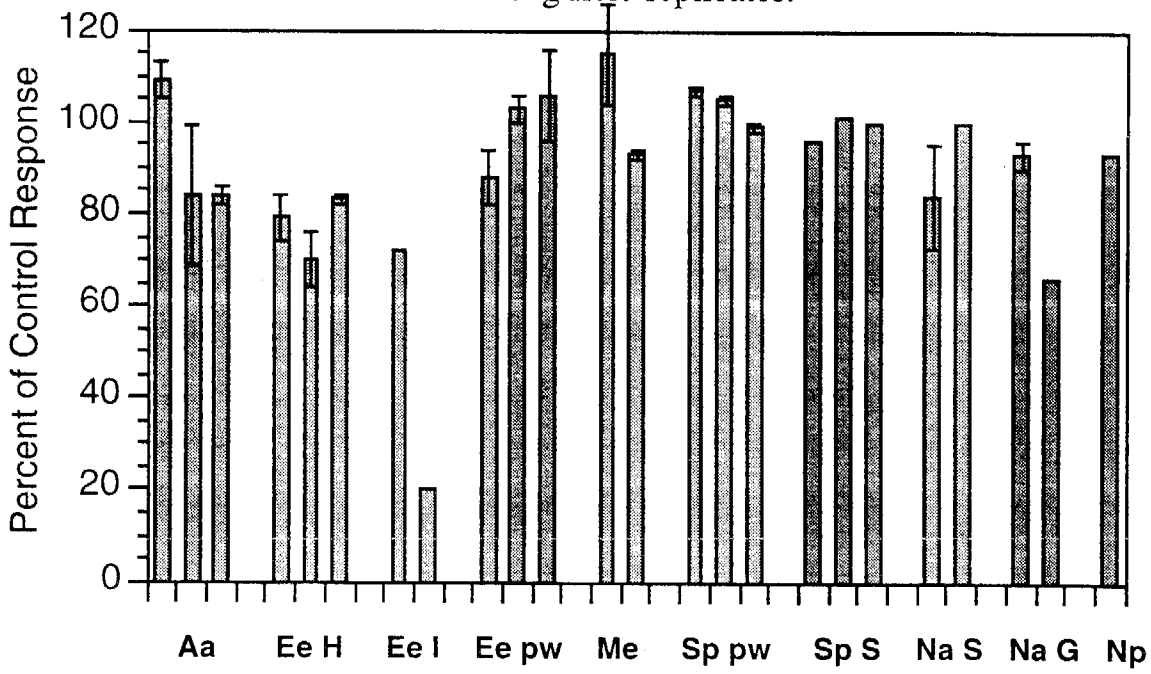


Figure 11. Results of toxicity tests of samples from Paradise Cove. Each column represents a sampling event; error bars are ± one standard deviation among field replicates.

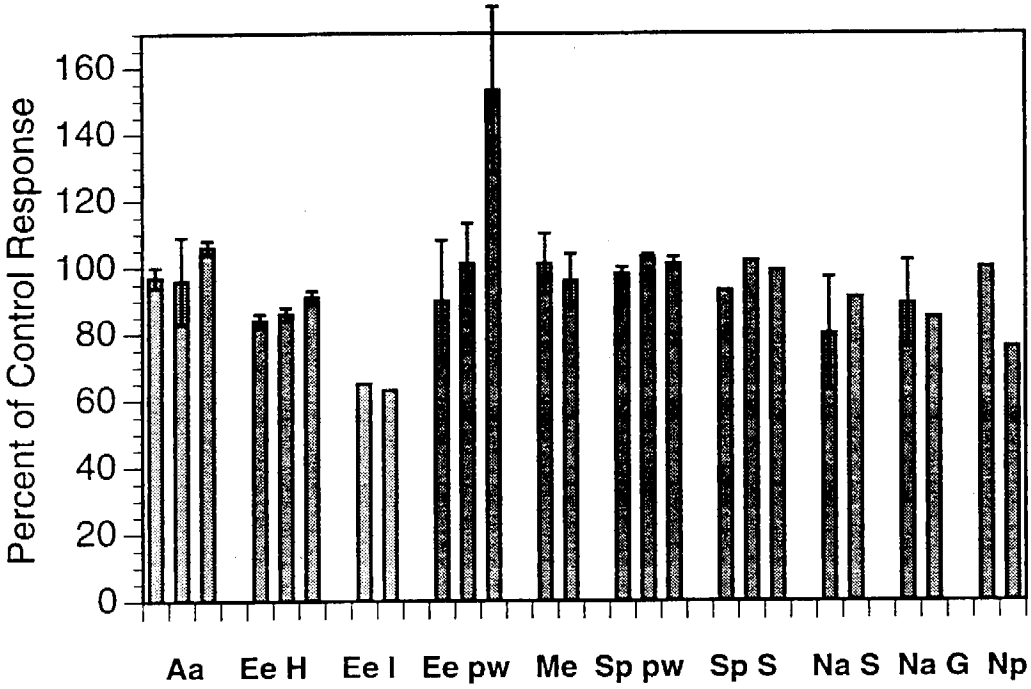


Figure 12. Results of toxicity tests of samples from the North South Bay site. Each column represents a sampling event; error bars are ± one standard deviation among field replicates.

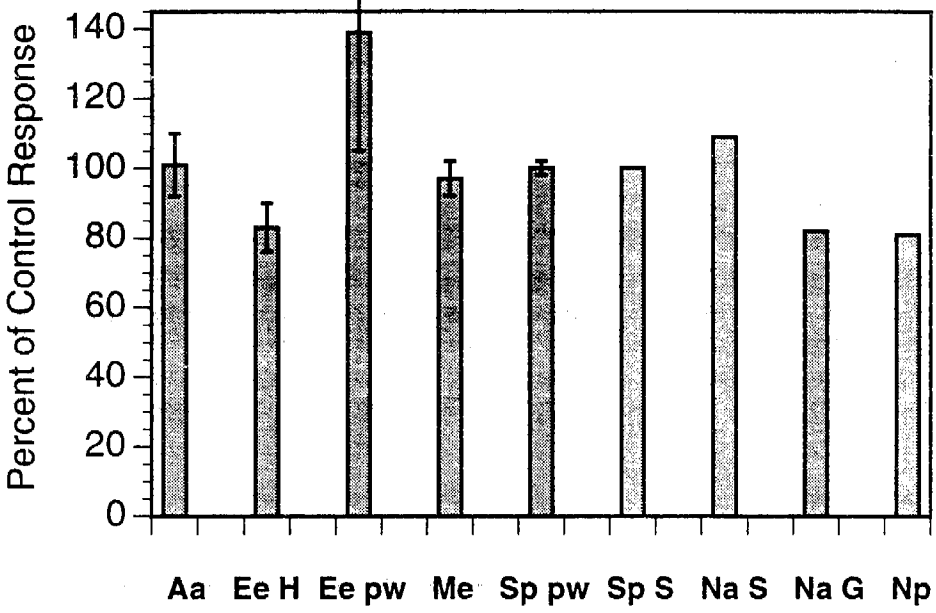


Figure 13. Results of toxicity tests of samples from the South South Bay site. Each column represents a sampling event; error bars are ± one standard deviation among field replicates.

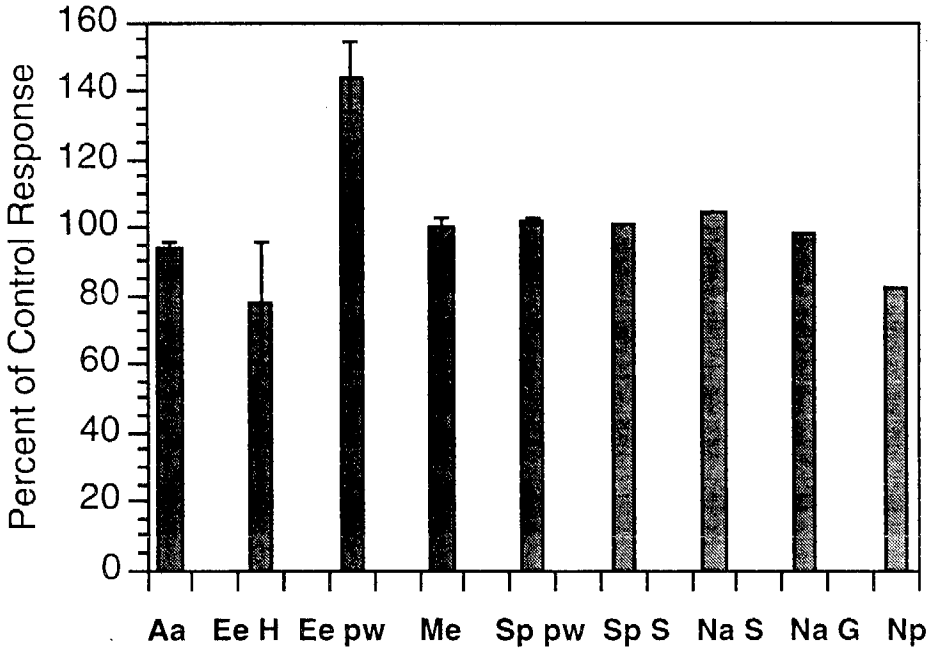


Figure 14. Results of toxicity tests of samples from Tomales Bay, Marconi Cove. Each column represents a sampling event; error bars are ± one standard deviation among field replicates.

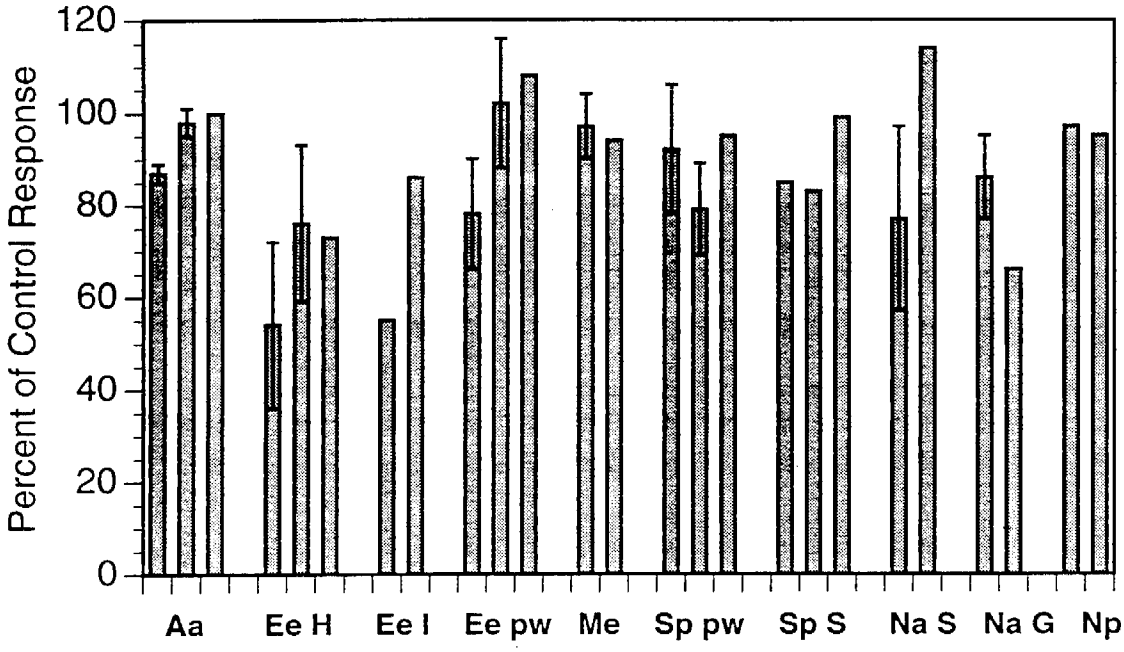


Figure 15. Results of toxicity tests of samples from Bolinas Lagoon. Each column represents a sampling event; error bars are ± one standard deviation among field replicates.

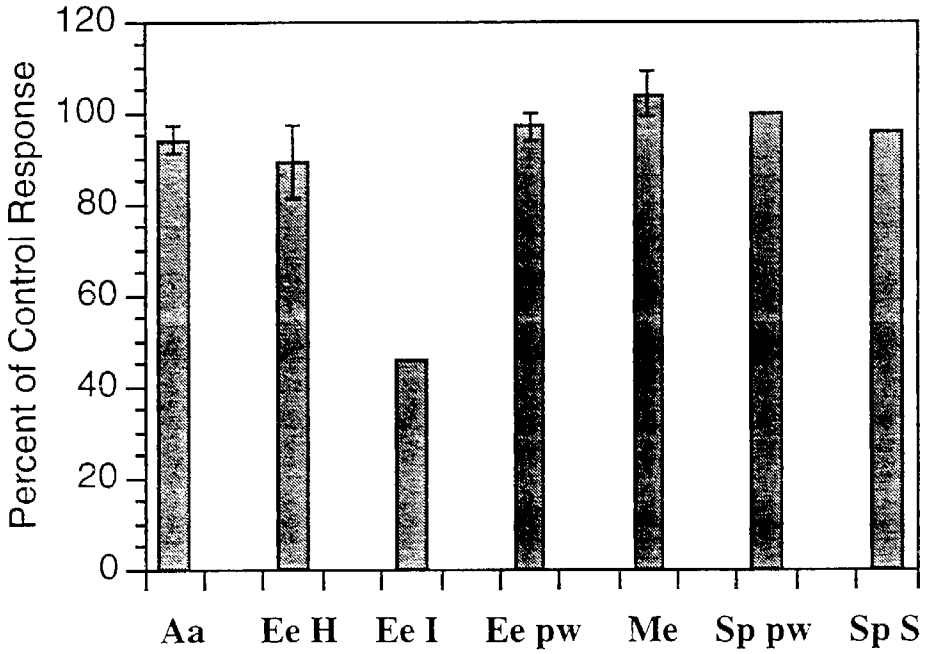
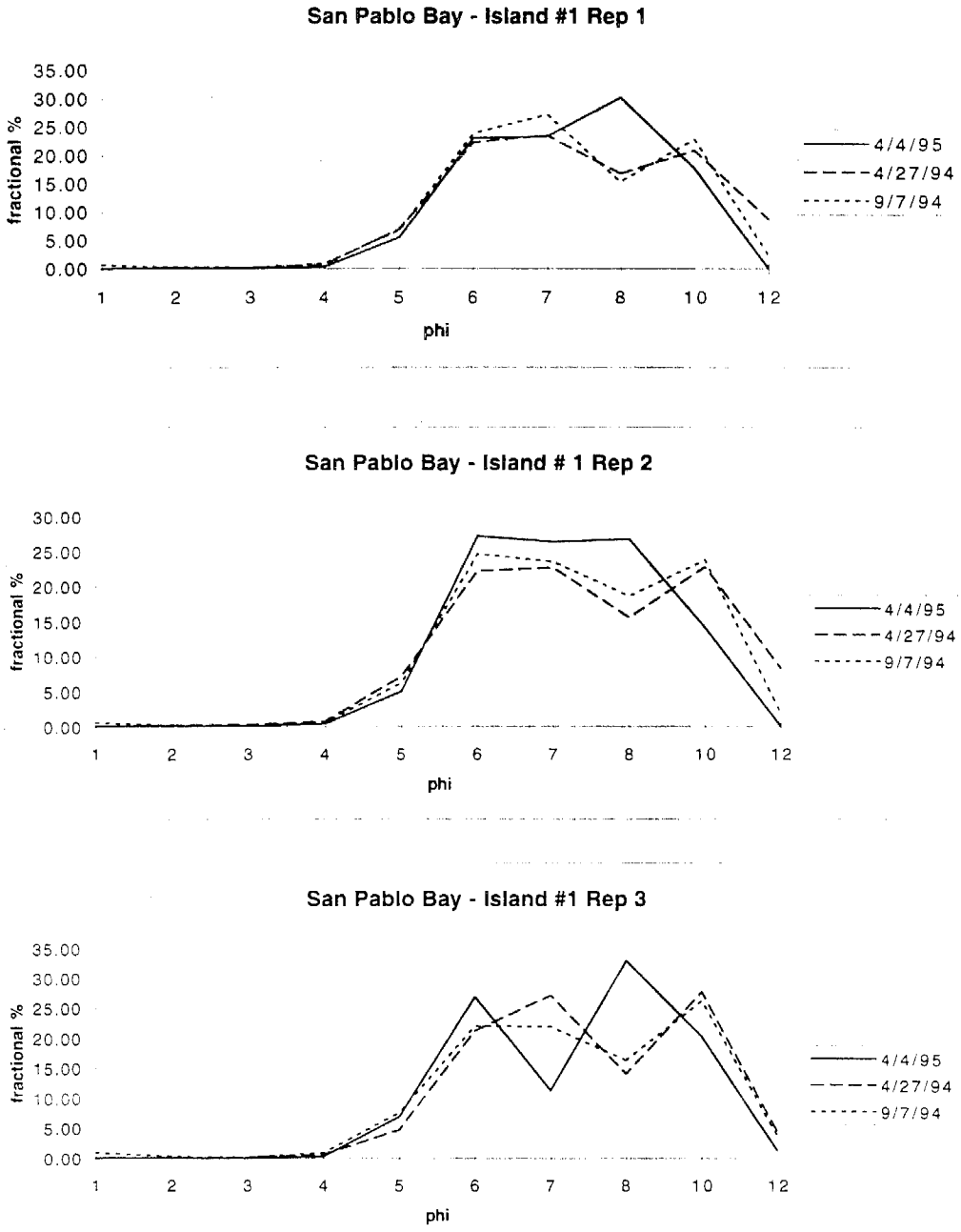


Figure 16. Grain size distribution in samples from San Pablo Bay, Island #1. Reps are field replicates, and each line represents a different sampling event.



Grain Size Equivalent Units for Figures 16 through 23:										
Phi Units	1	2	3	4	5	6	7	8	10	12
Micrometers	500	250	125	63	31	16	8	4	1	0.2
Description	Sand (> 63 µm)				(63 µm >) Silt (> 4 µm)				(4 >) Clay	

Figure 17. Grain size distribution in samples from San Pablo Bay, Tubbs Island. Reps are field replicates, and each line represents a different sampling event.

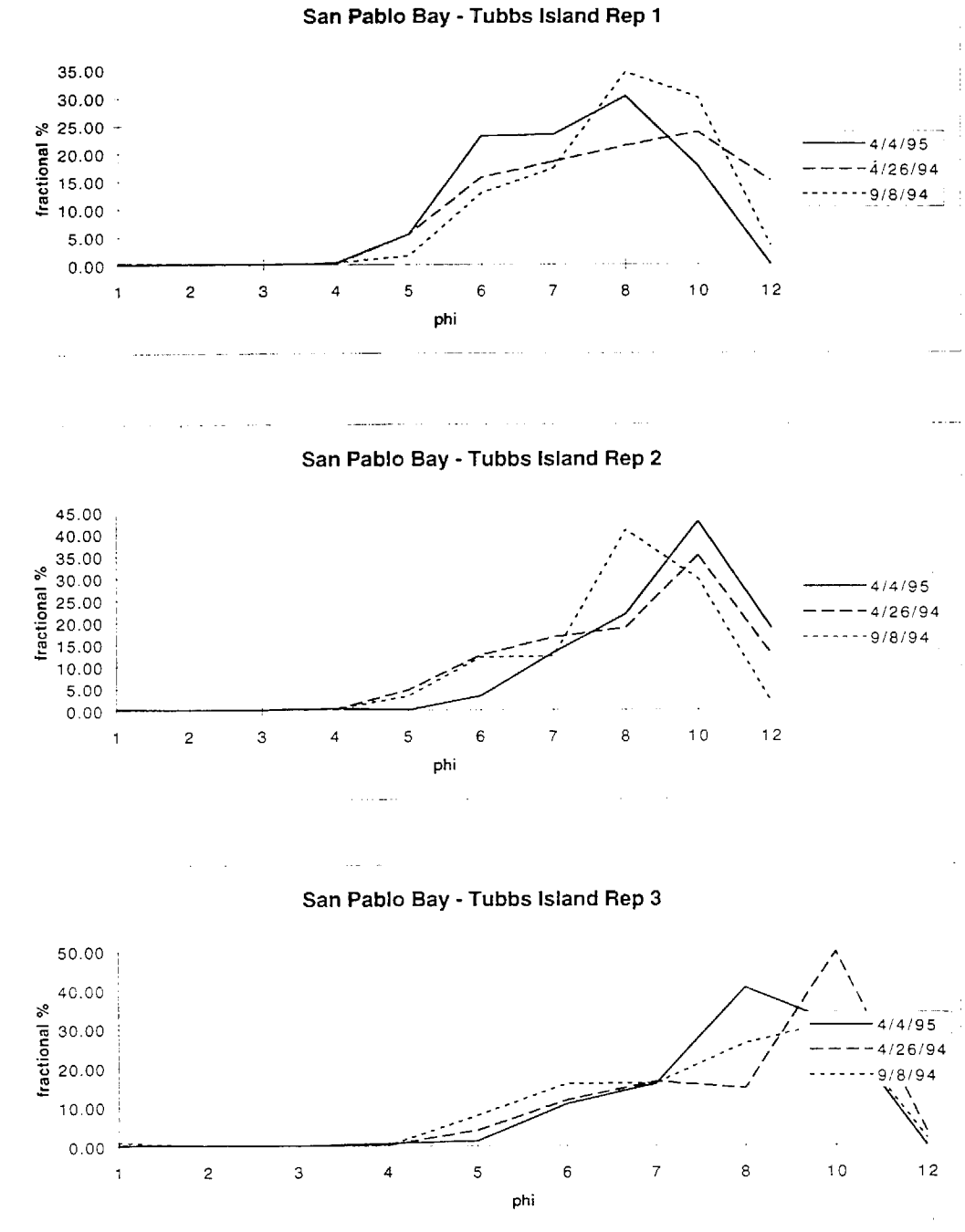


Figure 18. Grain size distribution in samples from Paradise Cove. Reps are field replicates, and each line represents a different sampling event.

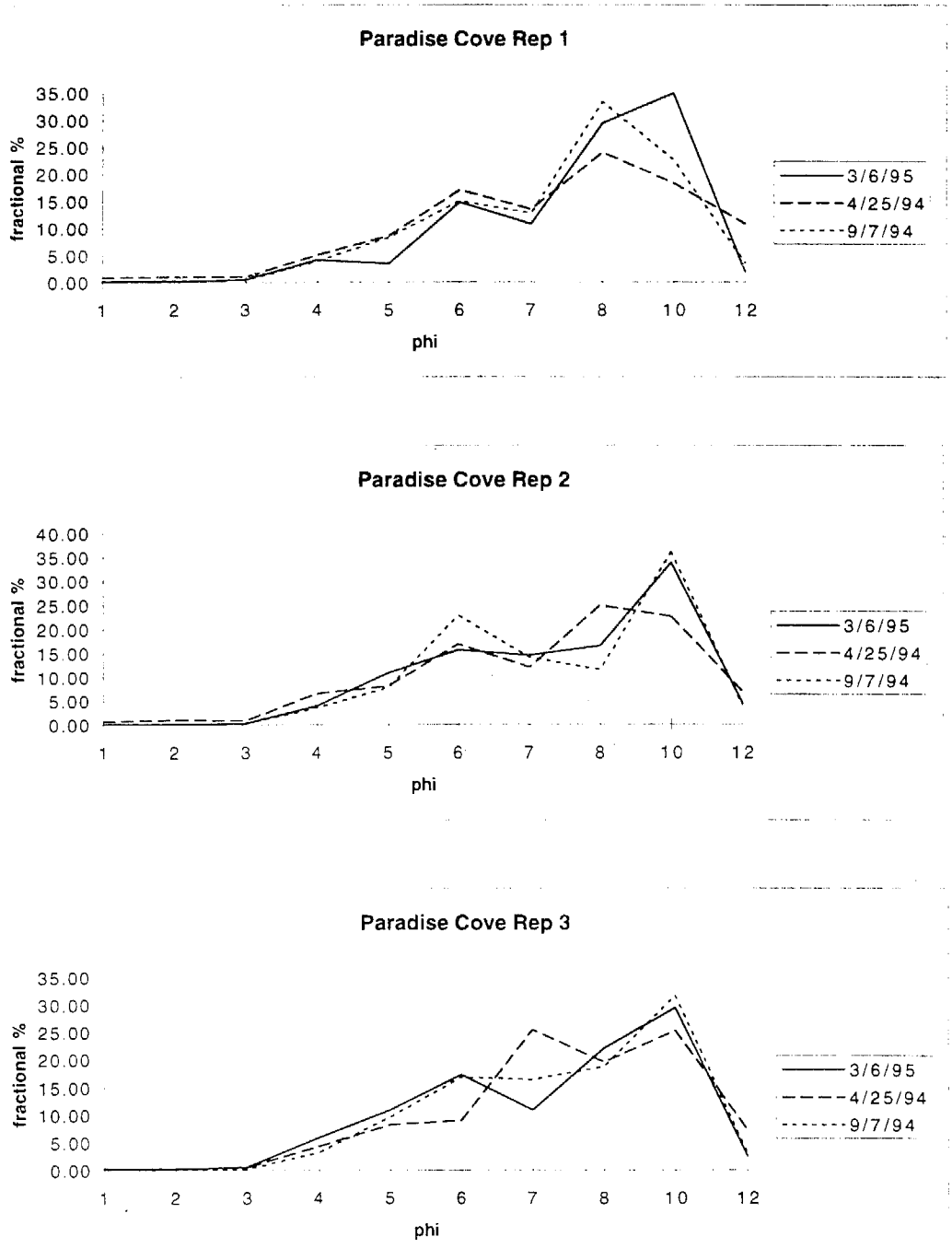
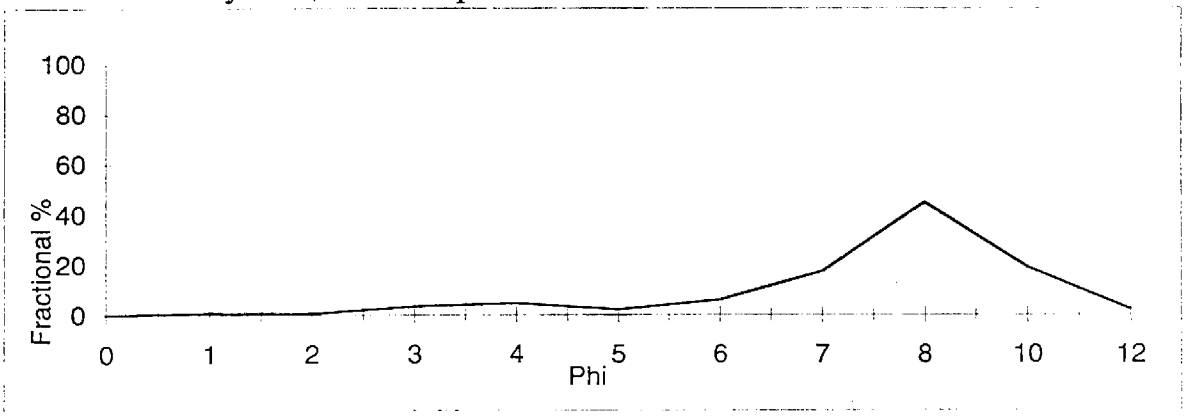
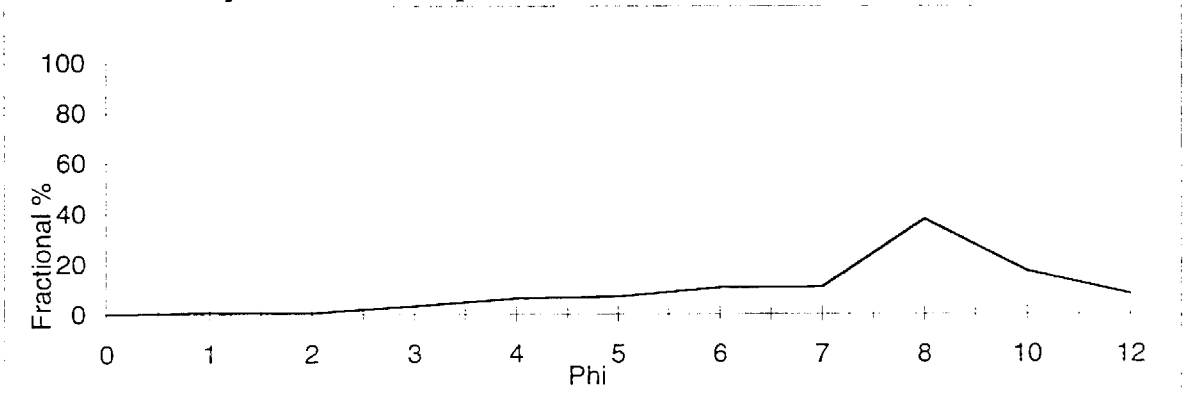


Figure 19. Grain size distribution at the North South Bay site.

North South Bay Site, Field Replicate 1, 03/06/95



North South Bay Site, Field Replicate 2, 03/06/95



North South Bay Site, Field Replicate 3, 03/06/95

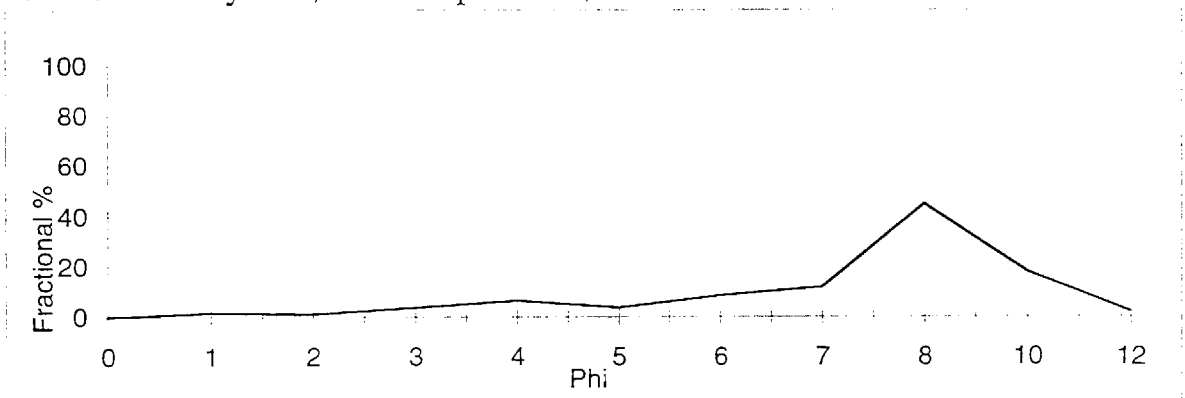
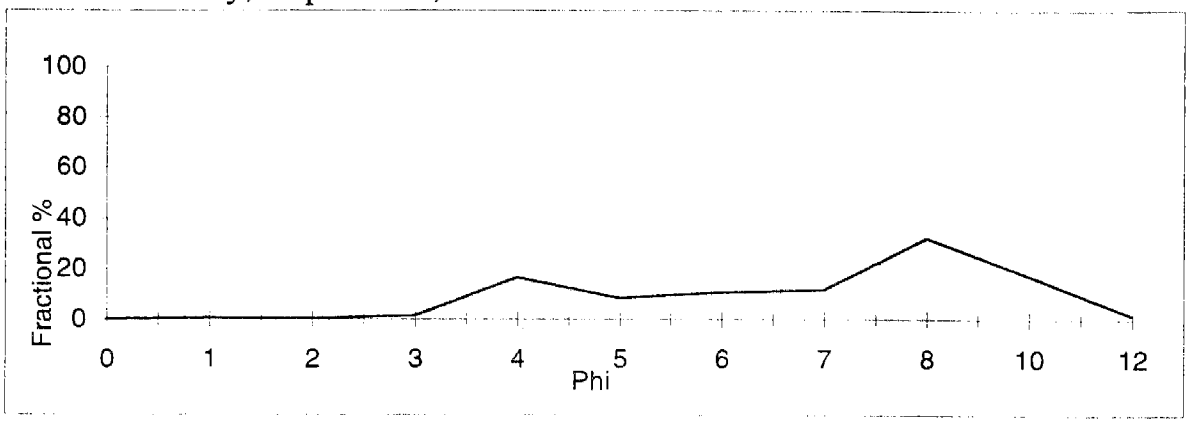
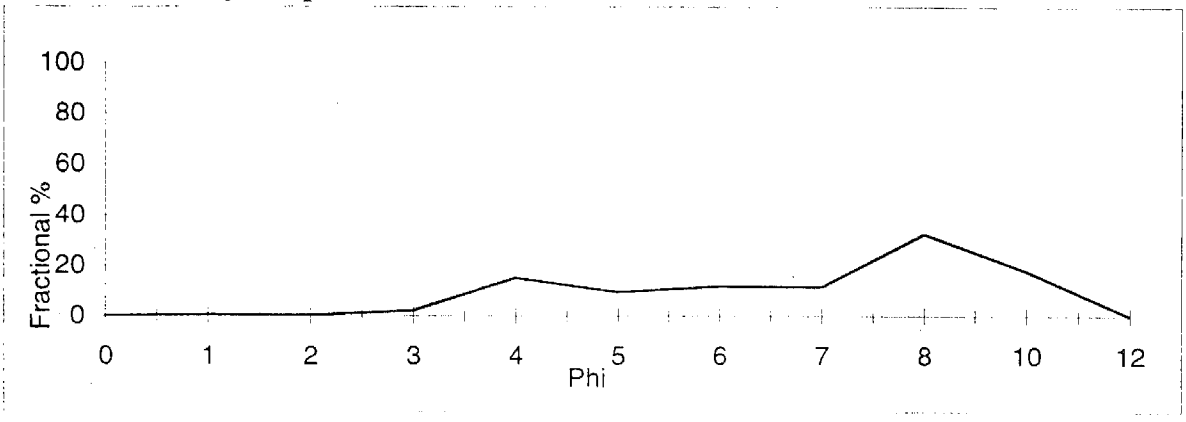


Figure 20. Grain size distribution at the South South Bay site.

South South Bay, Replicate 1, 03/07/95



South South Bay, Replicate 2, 03/07/95



South South Bay, Replicate 3, 03/07/95

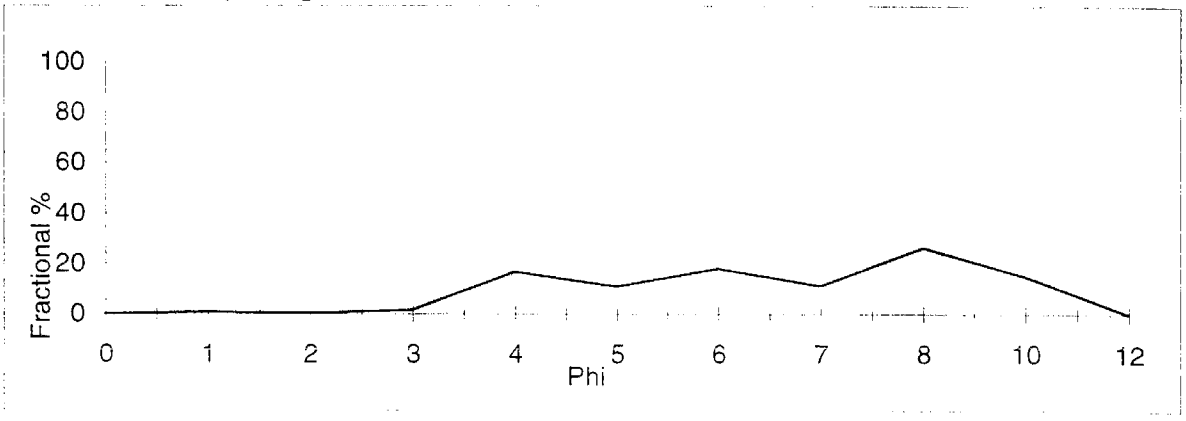


Figure 21. Grain size distribution in samples from Tomales Bay, Marconi Cove. Reps are field replicates, and each line represents a different sampling event.

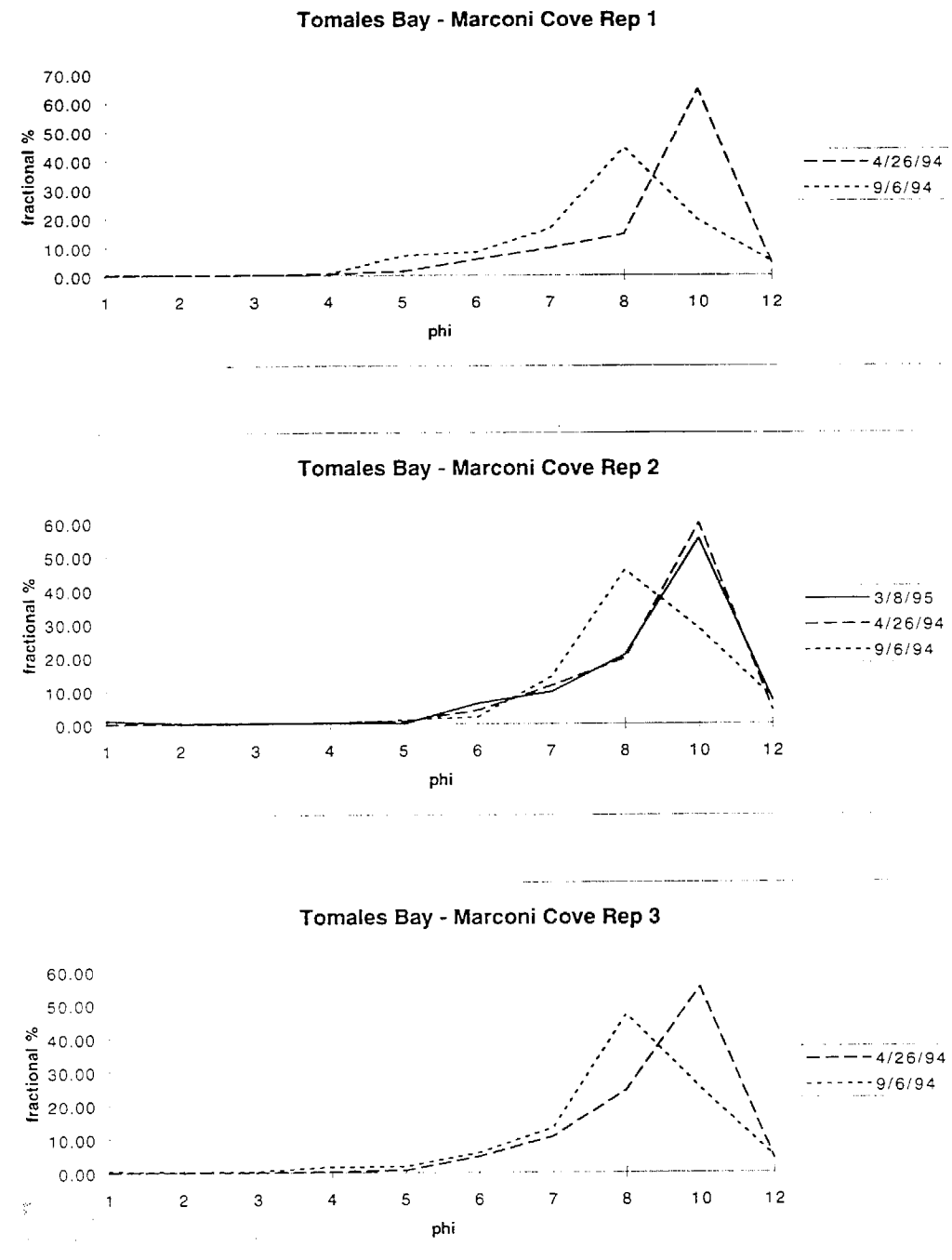
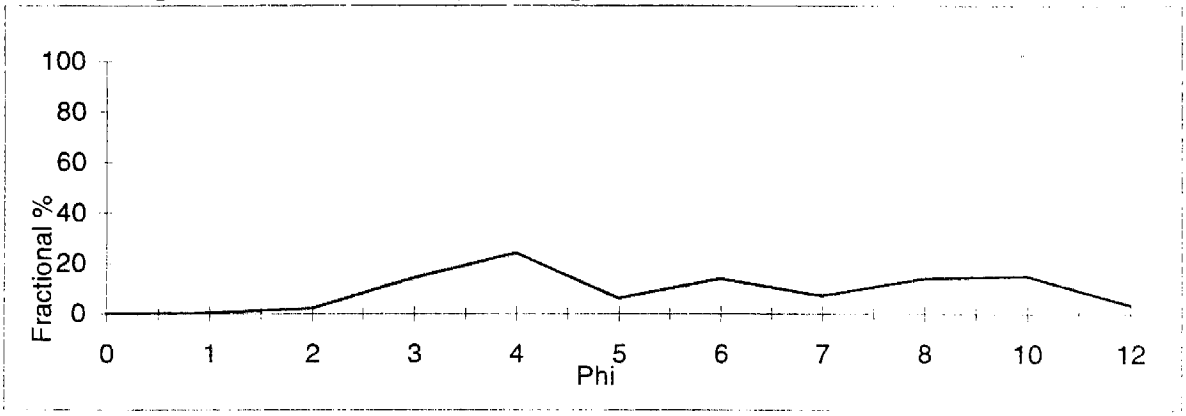
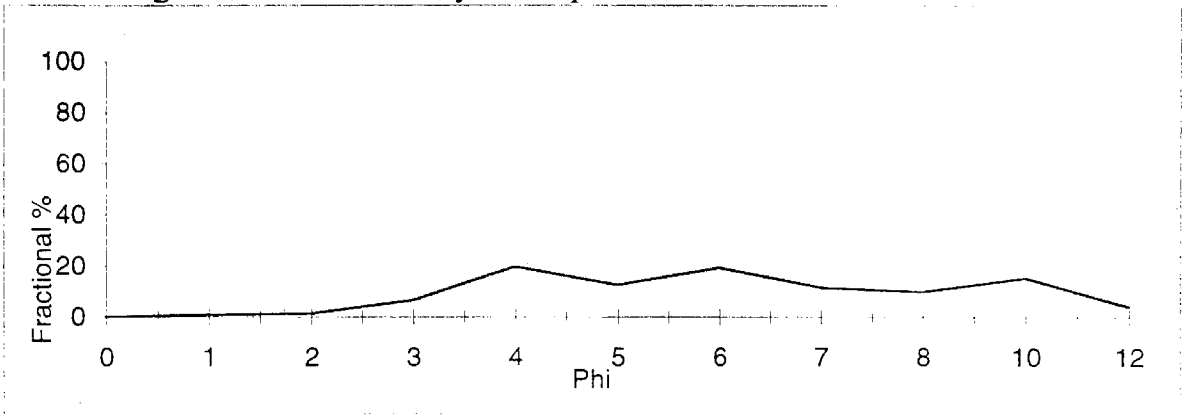


Figure 22. Grain size distribution in samples from Bolinas Lagoon.

Bolinas Lagoon, Audubon Canyon, Replicate 1, 4-25-94



Bolinas Lagoon, Audubon Canyon, Replicate 2, 4-25-94



Bolinas Lagoon, Audubon Canyon, Replicate 3, 4-25-94

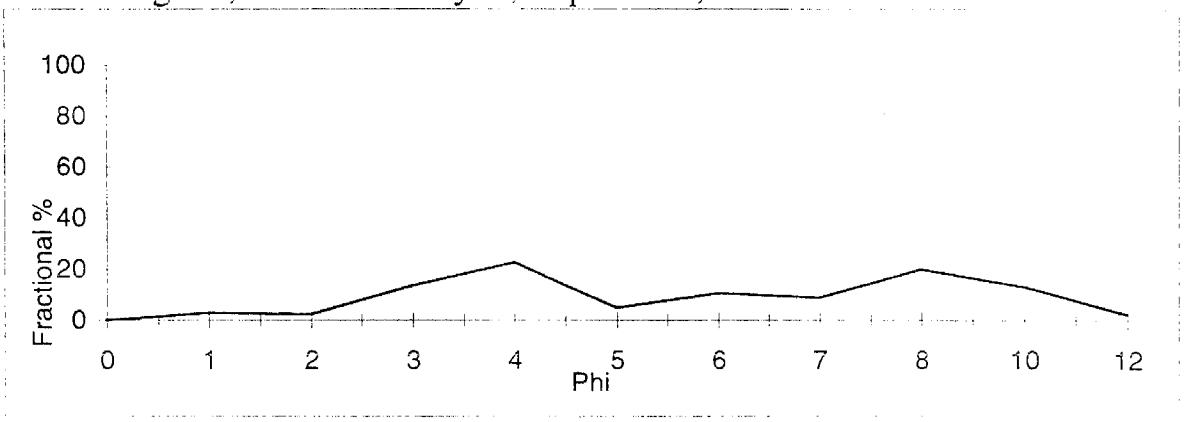
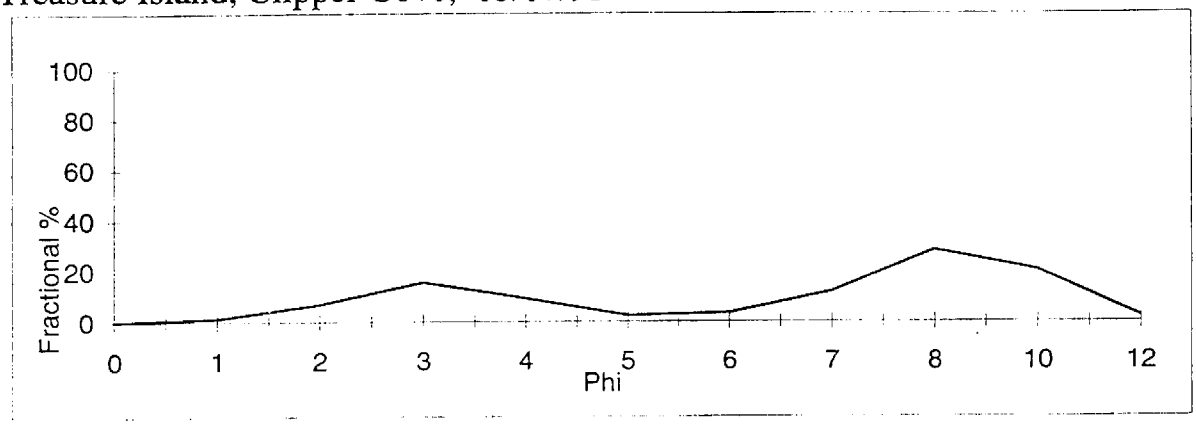
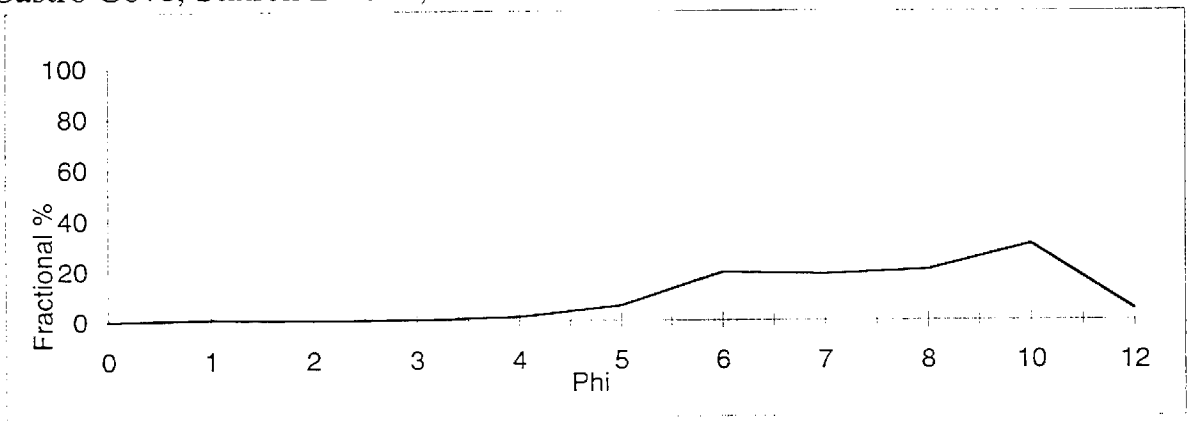


Figure 23. Grain size distribution in samples from test sites.

Treasure Island, Clipper Cove, 03/07/95



Castro Cove, Station EVS04, 09-05-94



Islais Creek, 09-05-94

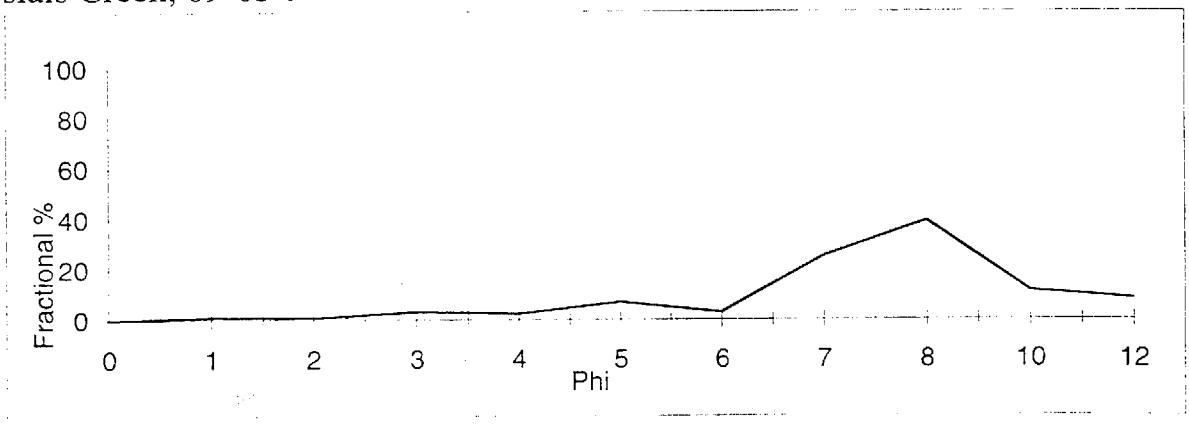


Figure 24a. Acceptability of control responses in solid-phase toxicity tests. Each bar represents the mean home sediment control response from tests conducted on different batches of samples. The control survival acceptability criterion is 90%.

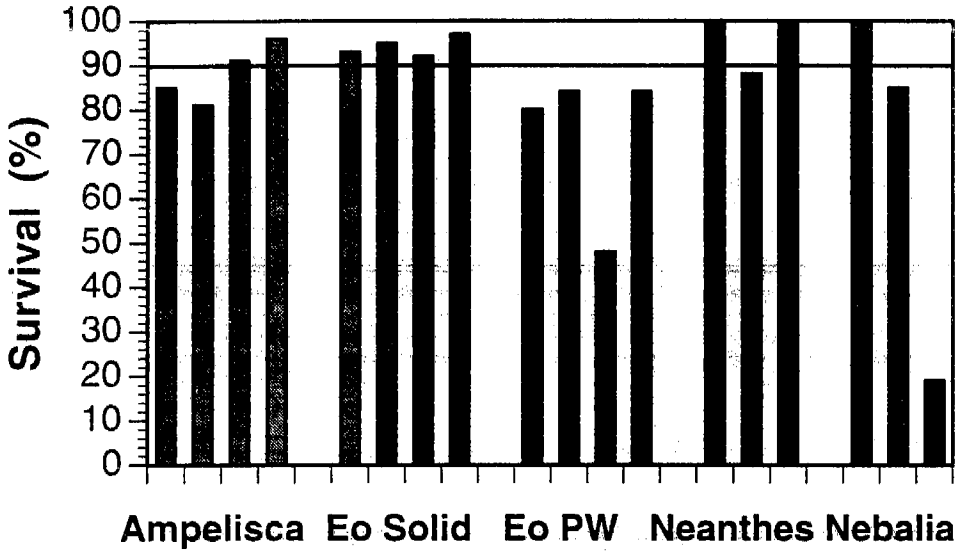


Figure 24b. Acceptability of control responses in larval toxicity tests of pore water and at the sediment-water interface. Bars represent mean control responses in tests from different batches of samples. The control acceptability criterion is 70%.

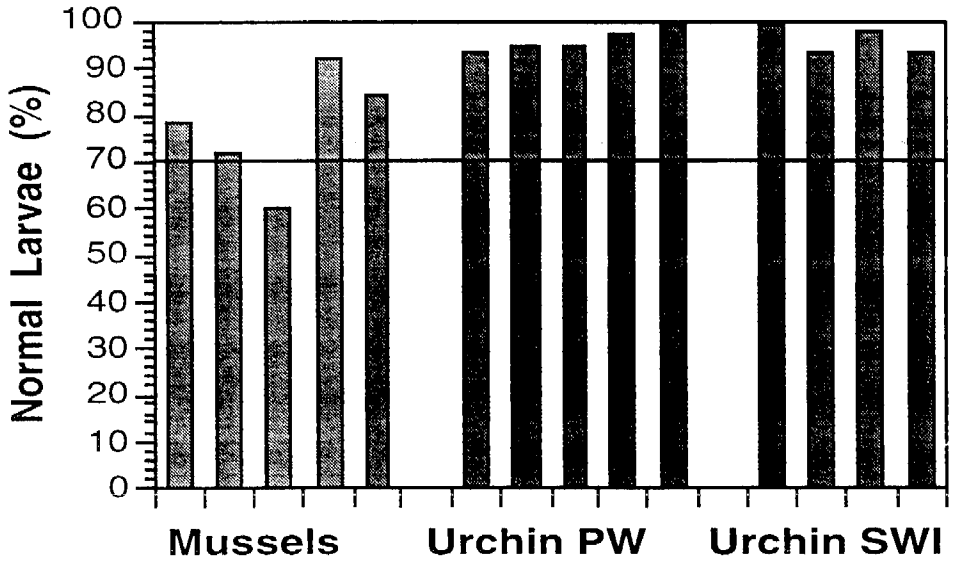


Figure 25. Variability among laboratory replicates. Bars represent average standard deviations (± sd) among five laboratory replicates for each test protocol. The number of samples tested ranged from 11 to 46, depending on the protocol.

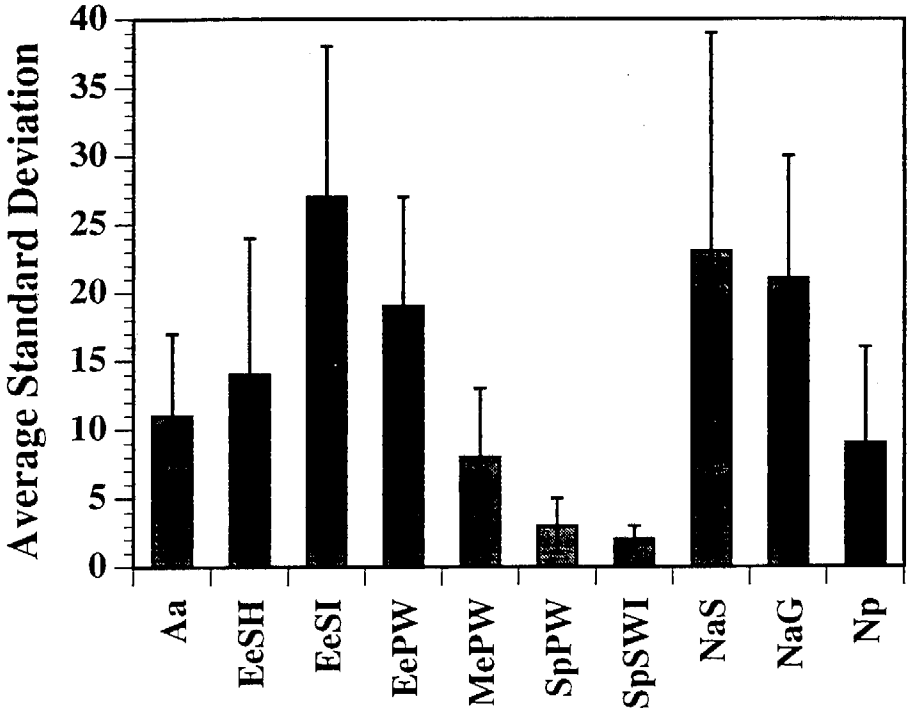


Figure 26. Comparison of test responses at a reference site (Paradise Cove), and two test sites (Castro Cove and Islais Creek). Error bars at Paradise Cove are ± one standard deviation among field replicates.

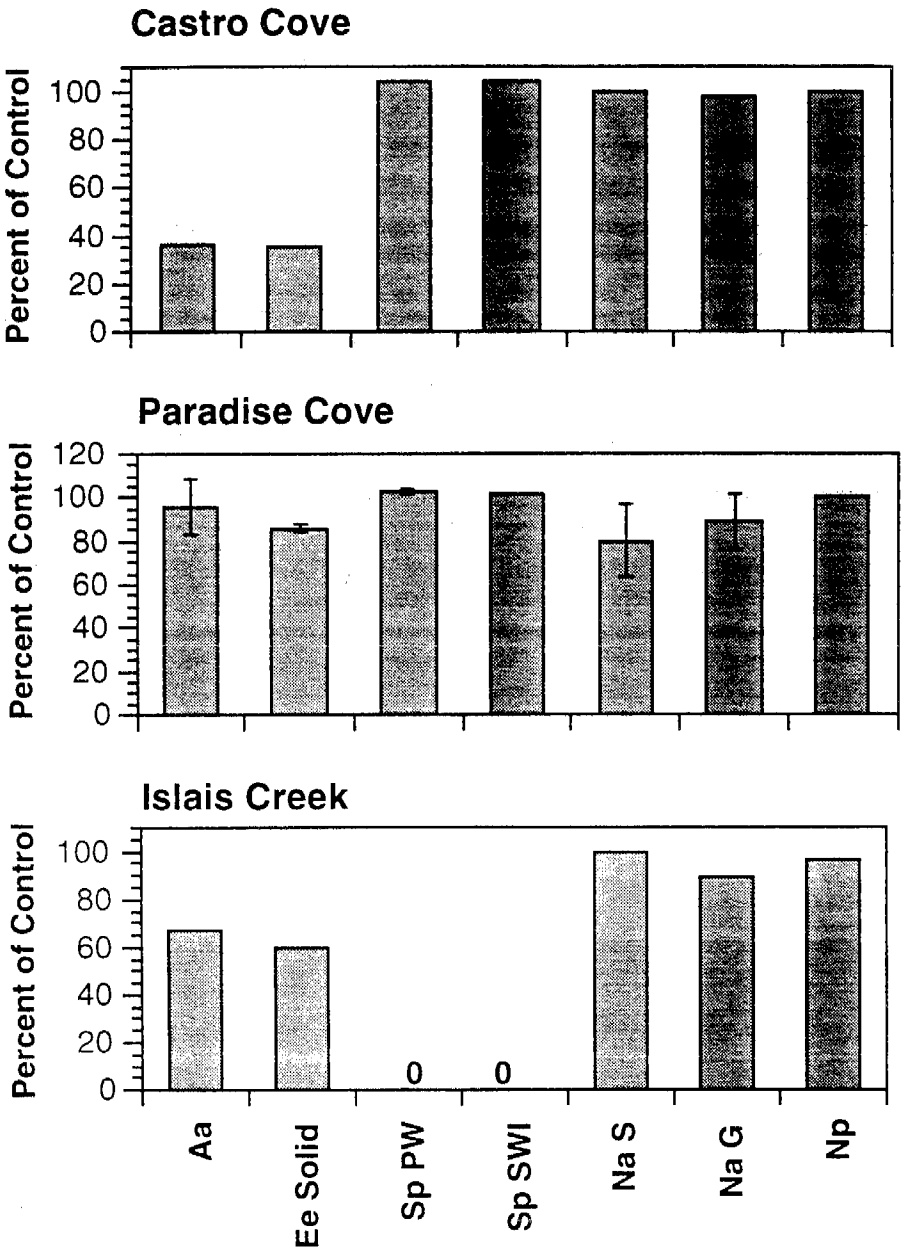


Figure 27a. Distribution of reference site data for the *Ampelisca* test in homogenized sediment. All data were from this study. There were no outliers identified or removed from this data set.

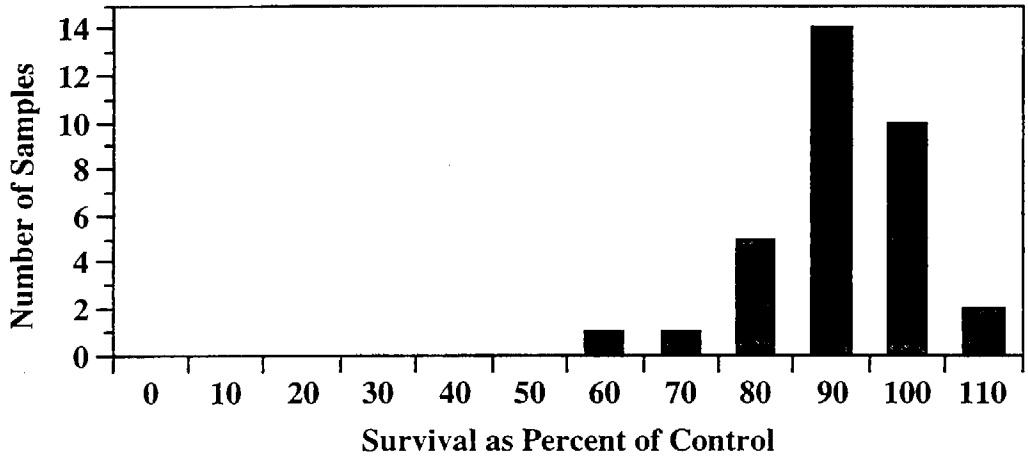


Figure 27b. Reference envelope tolerance limits for the *Ampelisca* test. All data were from this study. There were no outliers identified or removed from this data set.

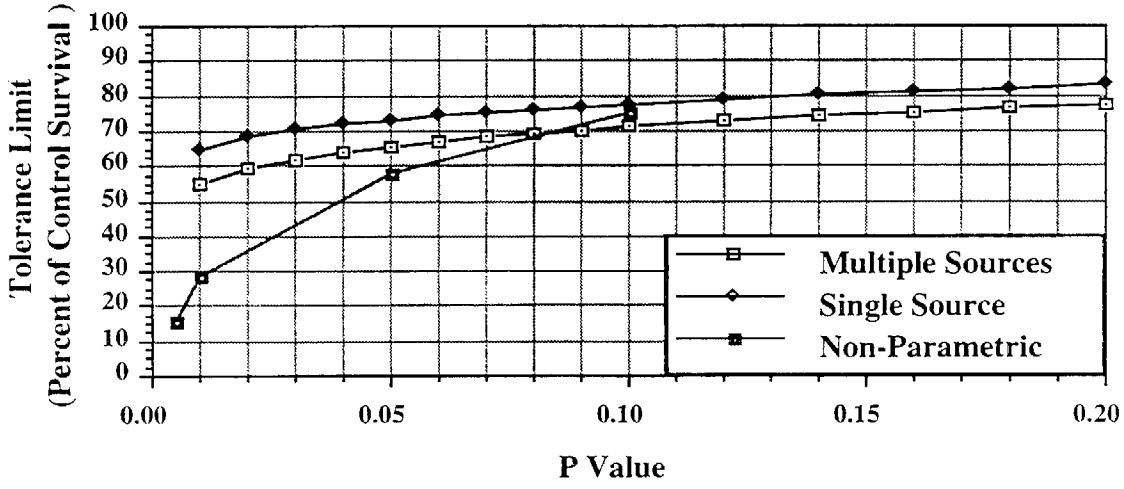


Figure 28a. Distribution of reference site data for the *Eohaustorius* test in homogenized sediment. Data are from this study. There were no outliers identified or removed from this data set.

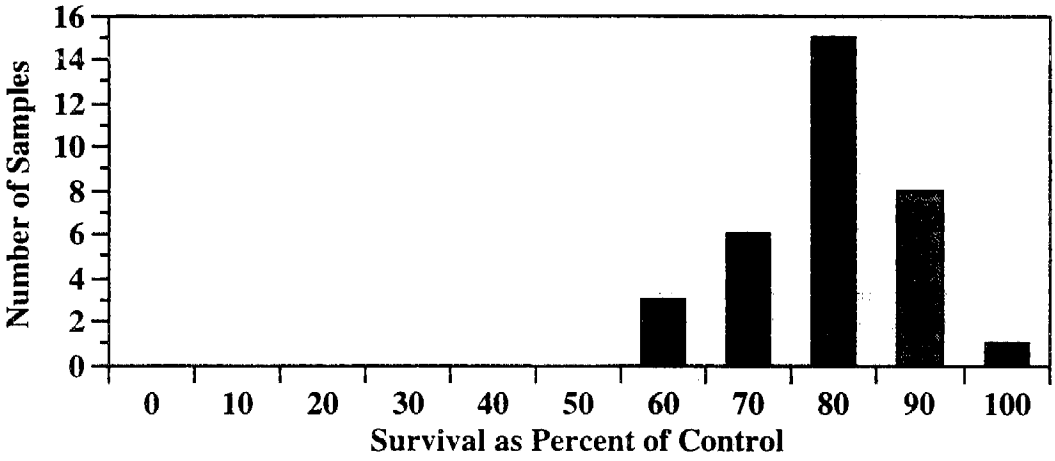


Figure 28b. Reference envelope tolerance limits for the *Eohaustorius* test in homogenized sediment. Data are from this study. There were no outliers identified or removed from this data set.

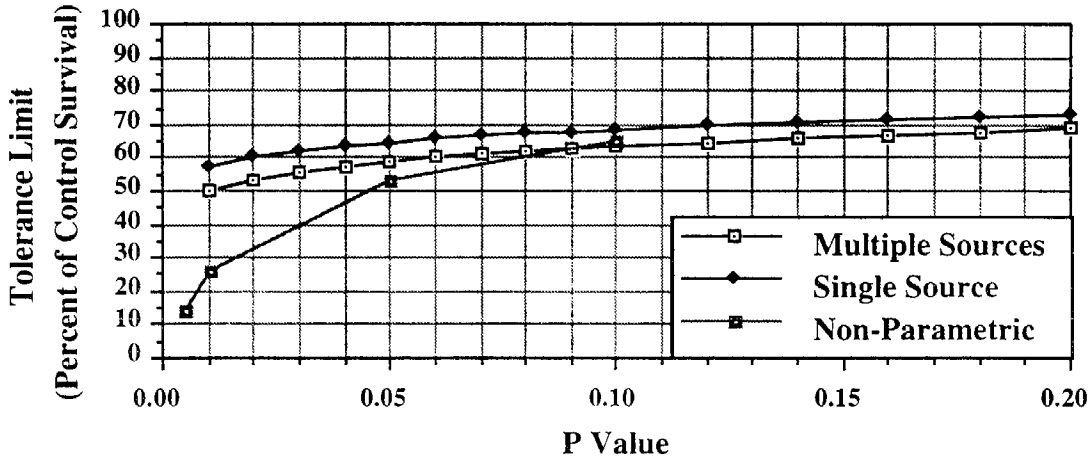


Figure 29a. Distribution of reference site data for the *Eohaustorius* test in homogenized sediment. Data are from this study plus additional BPTCP and RMP studies. There was one outlier identified, which is striped.

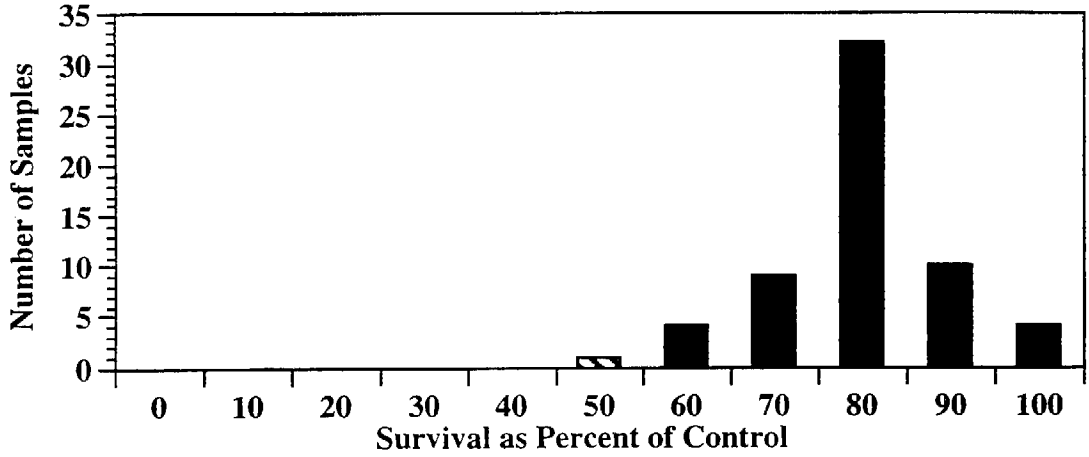


Figure 29b. Reference Envelope tolerance limits for the *Eohaustorius* test in homogenized sediment. Data are from this study plus additional BPTCP and RMP studies. The one outlier was removed for this analysis.

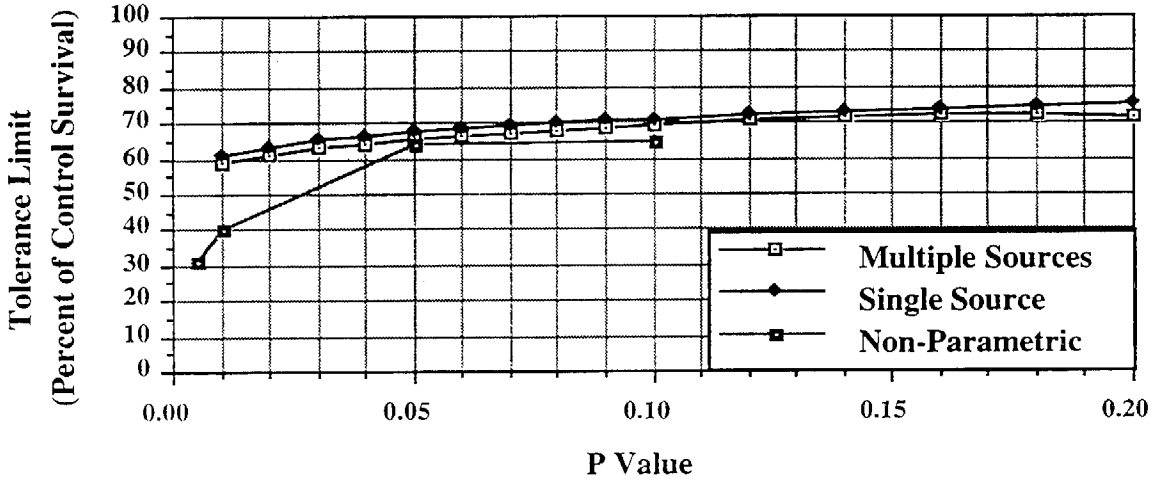


Figure 29c. Reference Envelope tolerance limits for the *Eohaustorius* test in homogenized sediment. Data are from this study plus additional BPTCP and RMP studies. The one outlier was retained for this analysis.

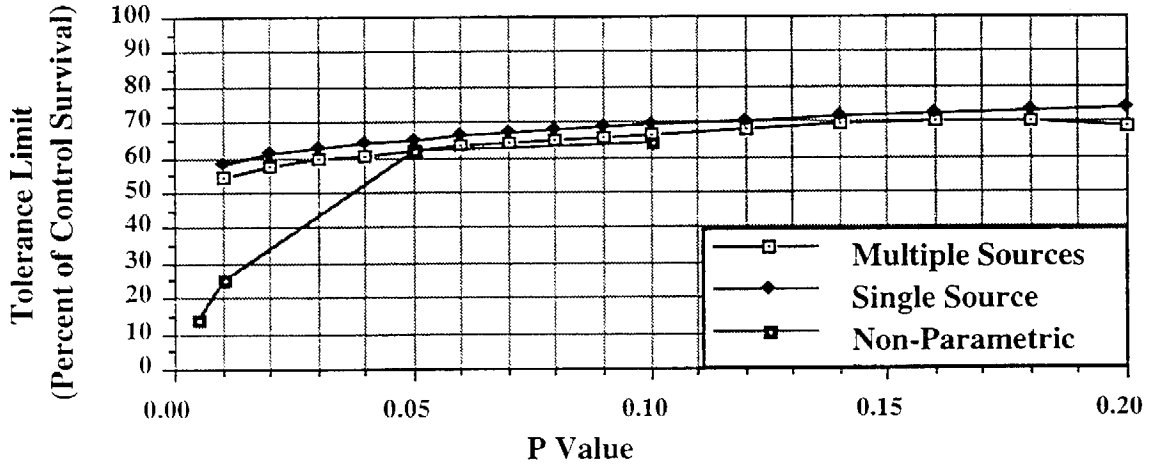


Figure 30a. Distribution of reference site data for the *Eohaustorius* test in intact sediment cores. Data are from this study. There was one outlier identified, which is striped.

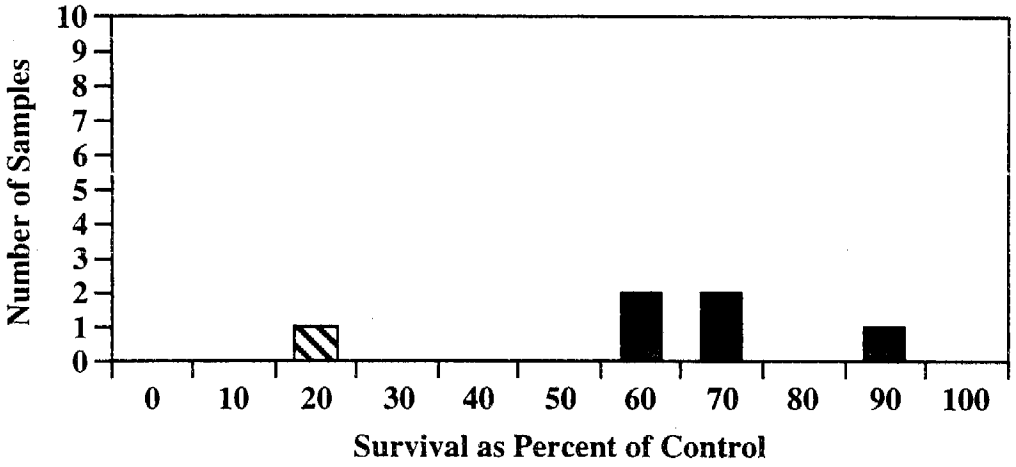


Figure 30b. Reference envelope tolerance limits for the *Eohaustorius* test in intact sediment cores. Data are from this study. There was one outlier identified, which was removed for this analysis.

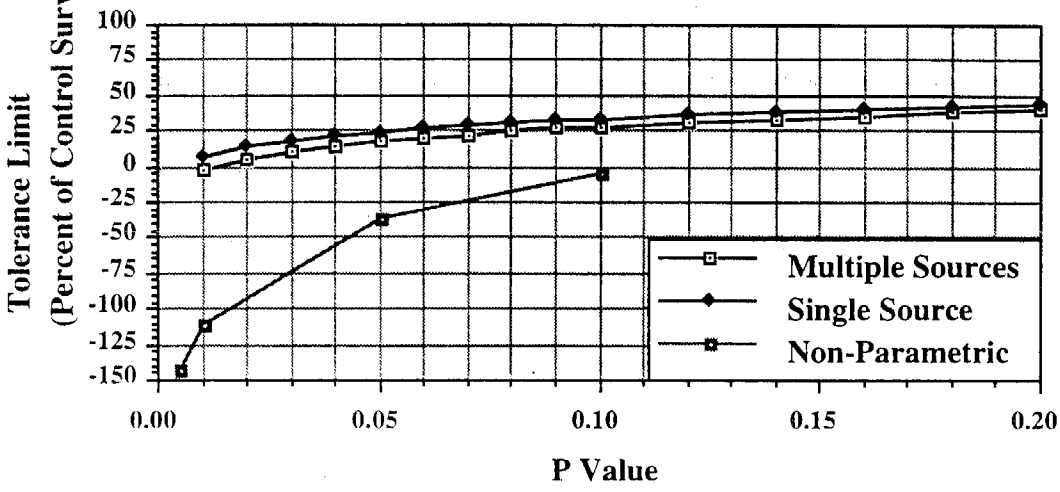


Figure 30c. Reference envelope tolerance limits for the *Eohaustorius* test in intact sediment cores. Data are from this study. There was one outlier identified, which was retained for this analysis.

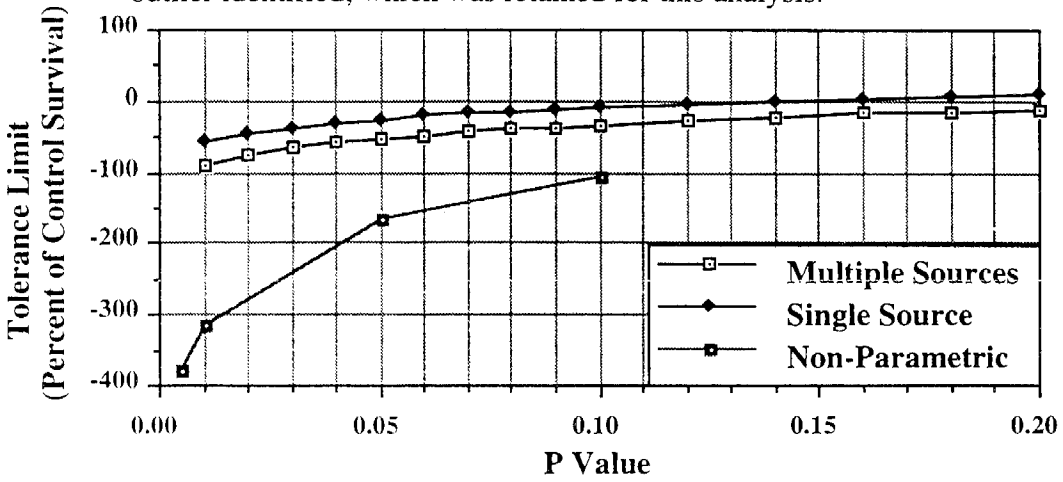


Figure 31a. Distribution of reference site data for the *Eohaustorius* test in sediment pore water. Data are from this study. There were two outliers identified, which are striped.

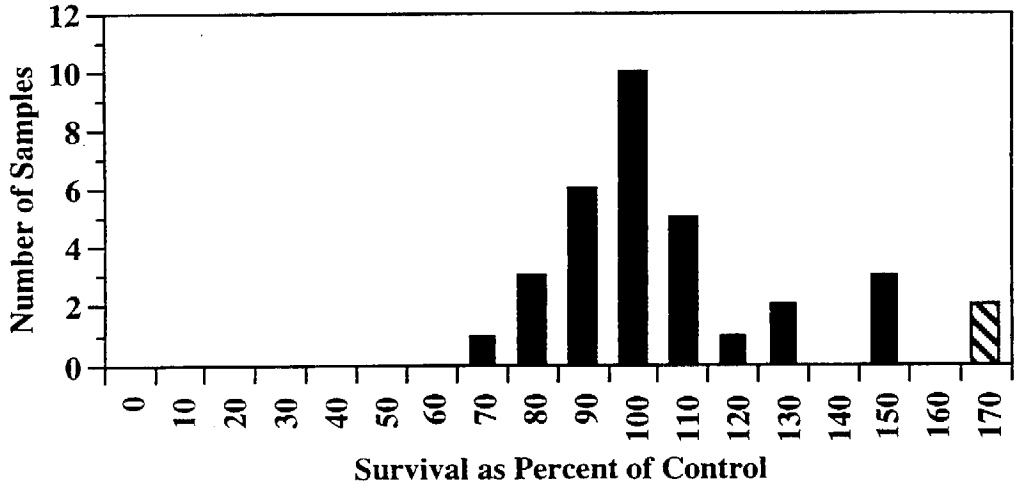


Figure 31b. Reference Envelope tolerance limits for the *Eohaustorius* test in sediment pore water. Data are from this study. The two outliers were removed for this analysis.

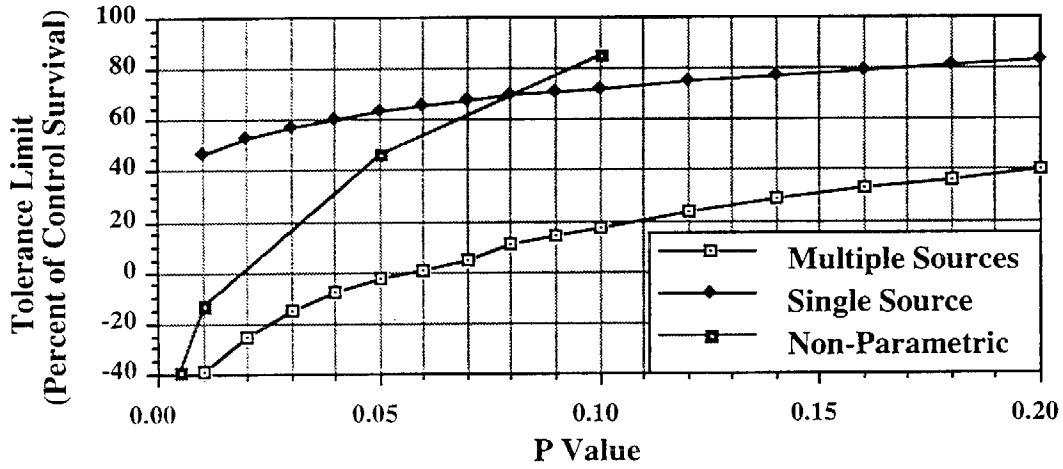


Figure 31c. Reference Envelope tolerance limits for the *Eohaustorius* test in sediment pore water. Data are from this study. The two outliers were retained for this analysis.

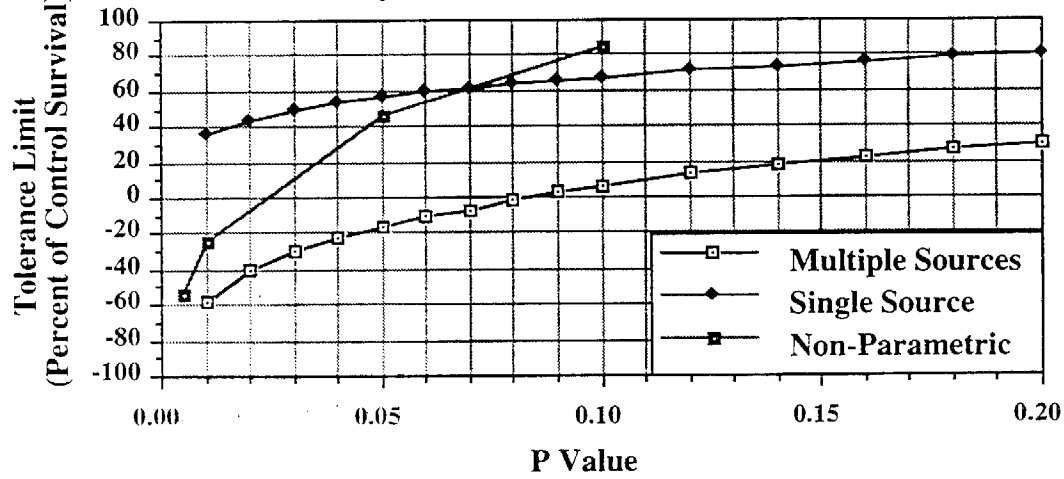


Figure 32a. Distribution of reference site data for the *Mytilus* test in sediment pore water. All data were from this study. There were no outliers identified or removed from this data set.

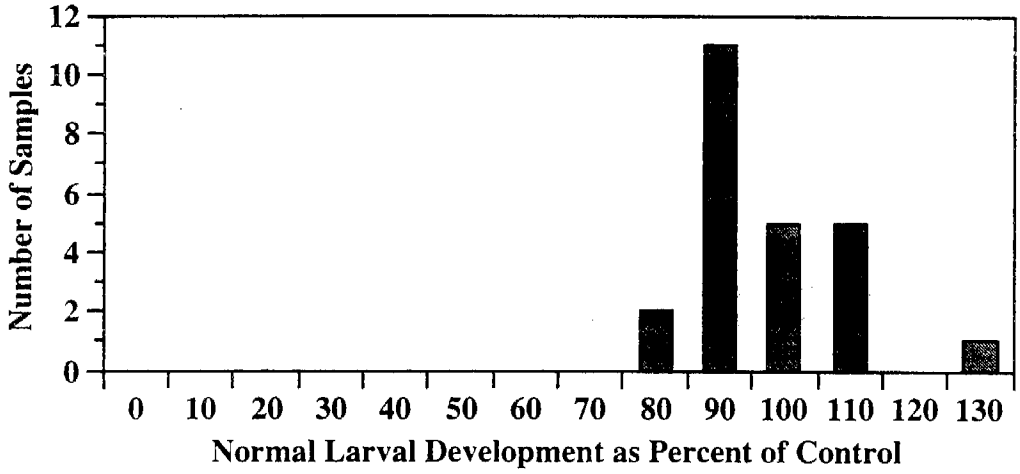


Figure 32b. Reference envelope tolerance limits for the *Mytilus* test. All data were from this study. There were no outliers identified or removed from this data set.

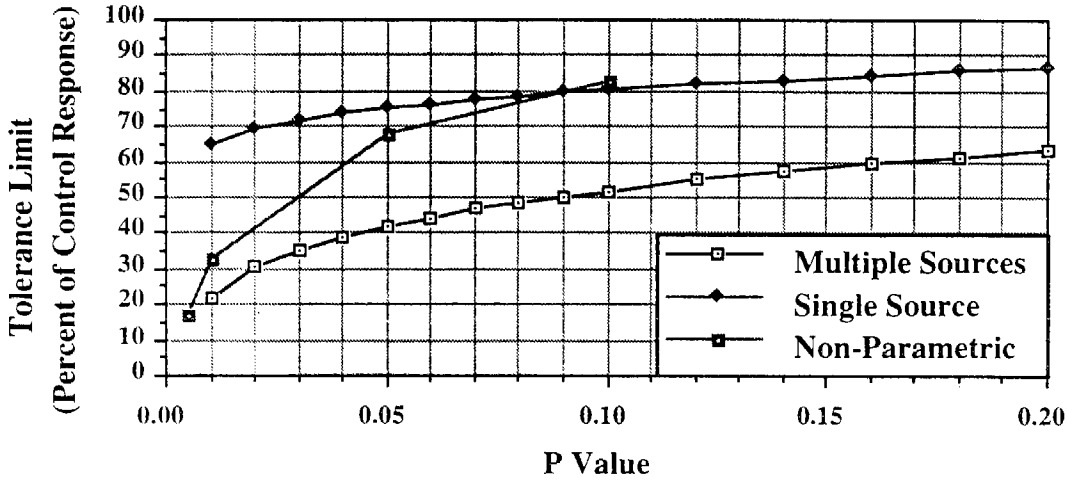


Figure 33a. Distribution of reference site data for the sea urchin test in sediment pore water. All data were from this study. There were no outliers identified or removed from this data set.

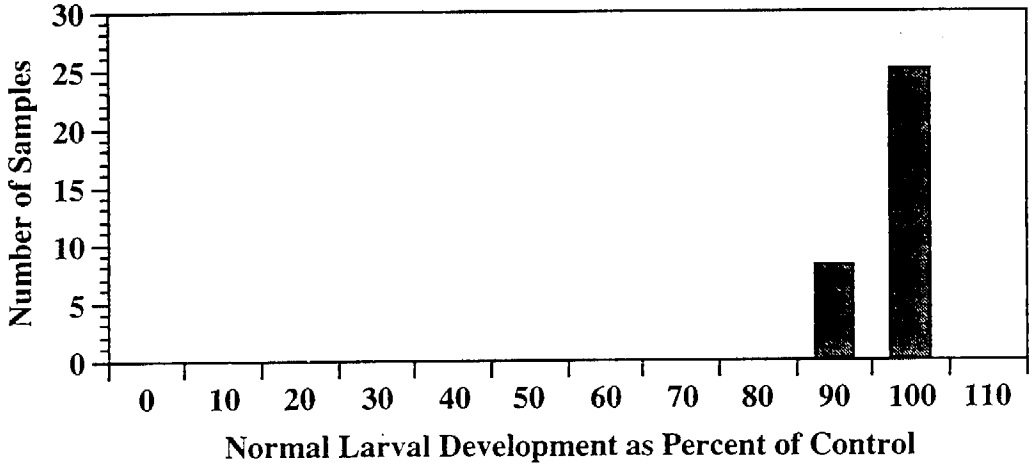


Figure 33b. Reference envelope tolerance limits for the sea urchin test. All data were from this study. There were no outliers identified or removed from this data set.

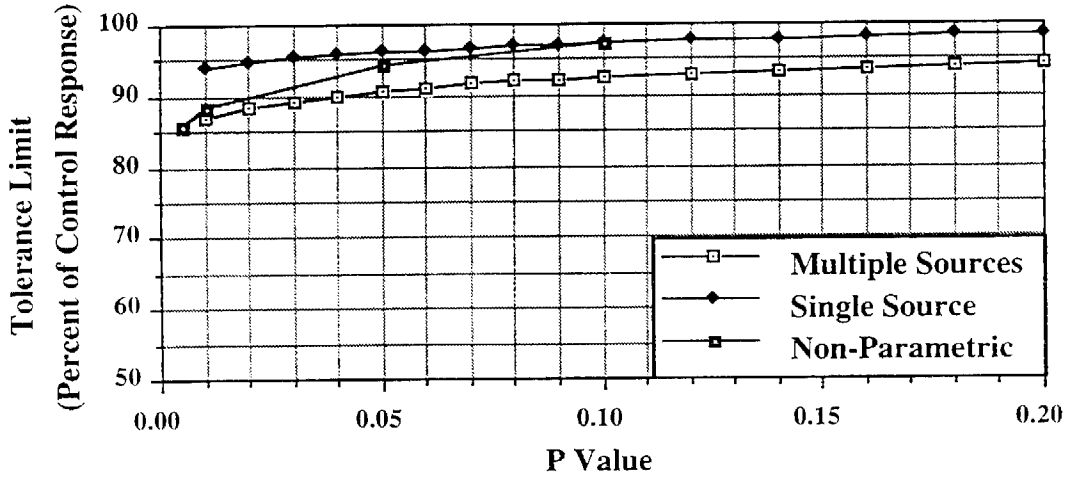


Figure 34a. Distribution of reference site data for the sea urchin test in sediment pore water. Data are from this study plus additional BPTCP and RMP studies. There were four outliers identified, which are striped.

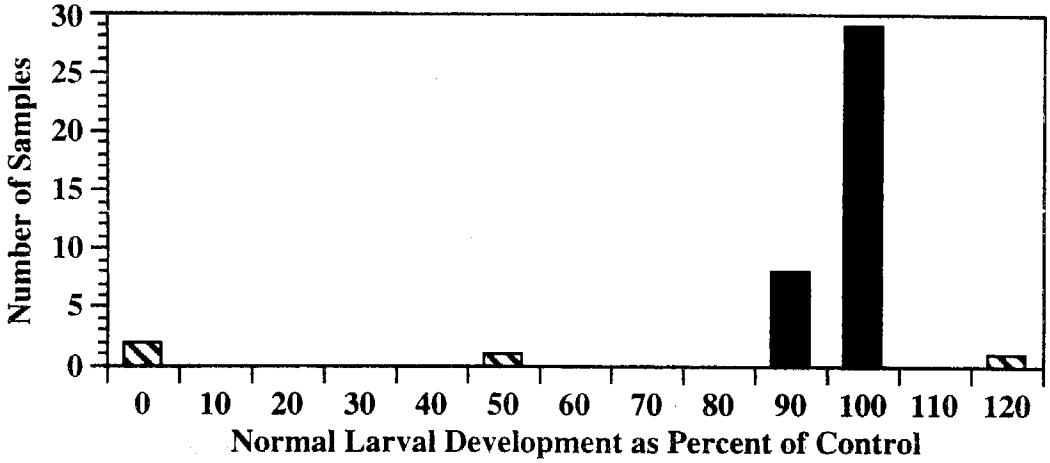


Figure 34b. Reference Envelope tolerance limits for the sea urchin test in sediment pore water. Data are from this study plus additional BPTCP and RMP studies. There were four outliers identified, which were removed for this analysis.

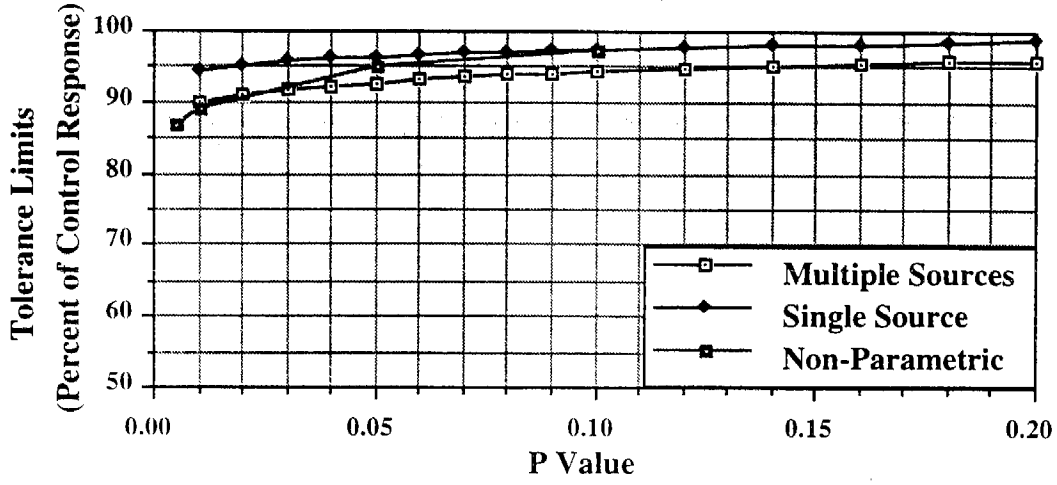


Figure 34c. Reference Envelope tolerance limits for the sea urchin test in sediment pore water. Data are from this study plus additional BPTCP and RMP studies. There were four outliers identified, which were retained for this analysis.

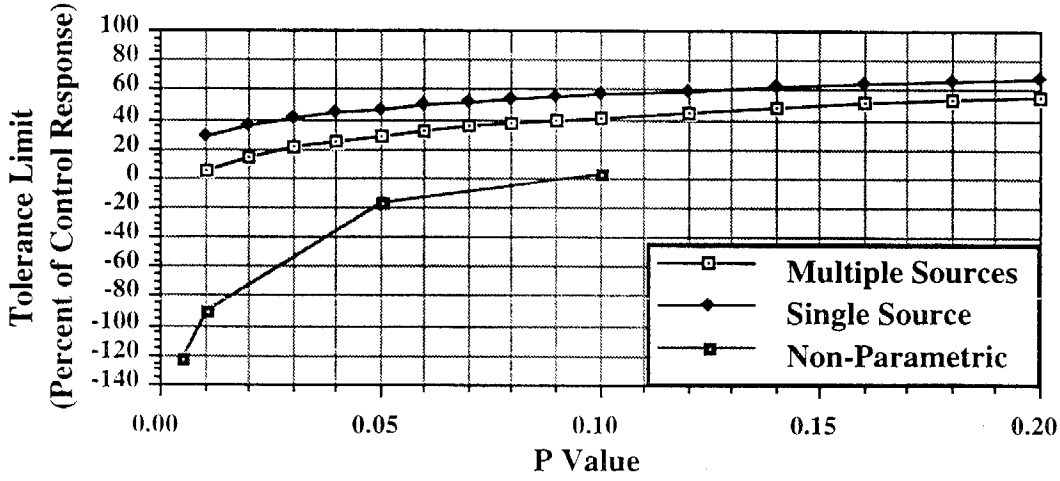


Figure 35a. Distribution of reference site data for the sea urchin test at the sediment/water interface. All data were from this study. There were no outliers identified or removed from this data set.

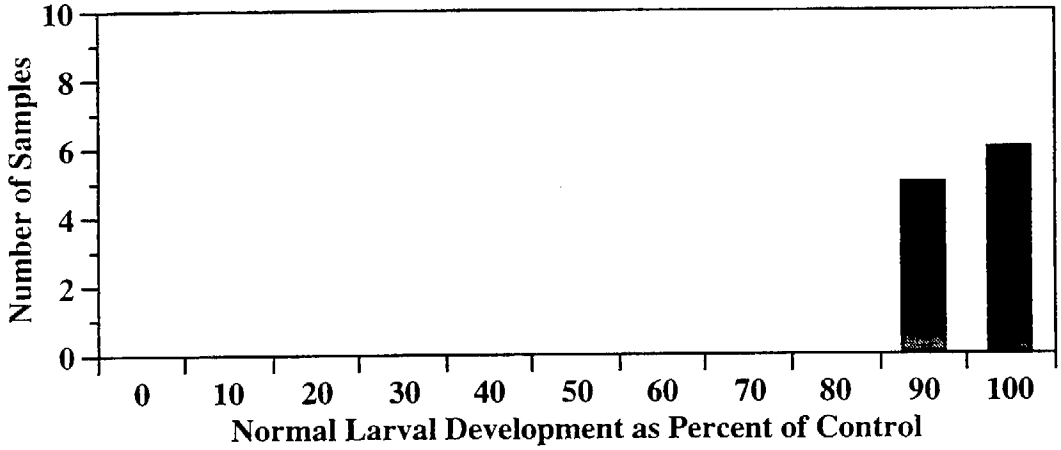


Figure 35b. Reference envelope tolerance limits for the sea urchin test at the sediment/water interface. There were no outliers identified or removed from this data set.

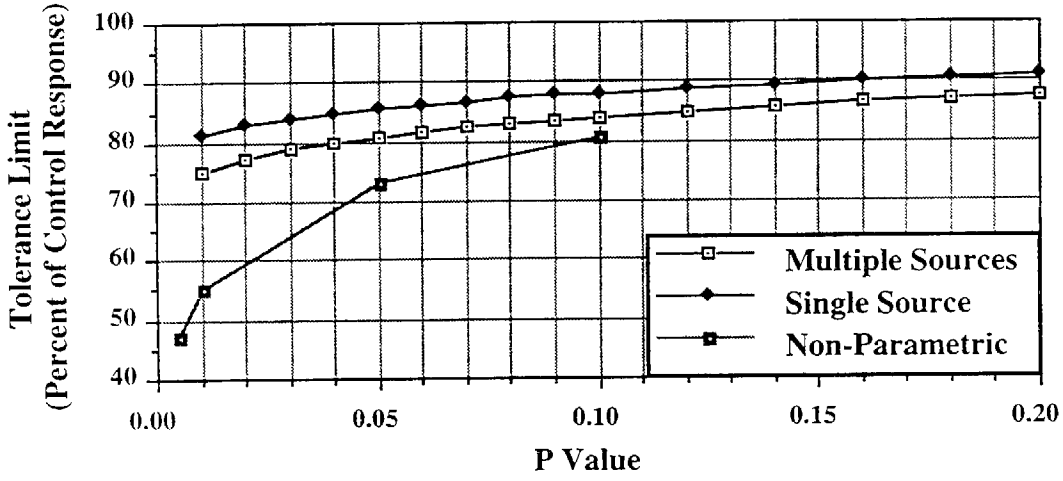


Figure 36a. Distribution of reference site data for the sea urchin test at the sediment/water interface. Data are from this study plus additional BPTCP and RMP studies. There were no outliers identified or removed from this data set.

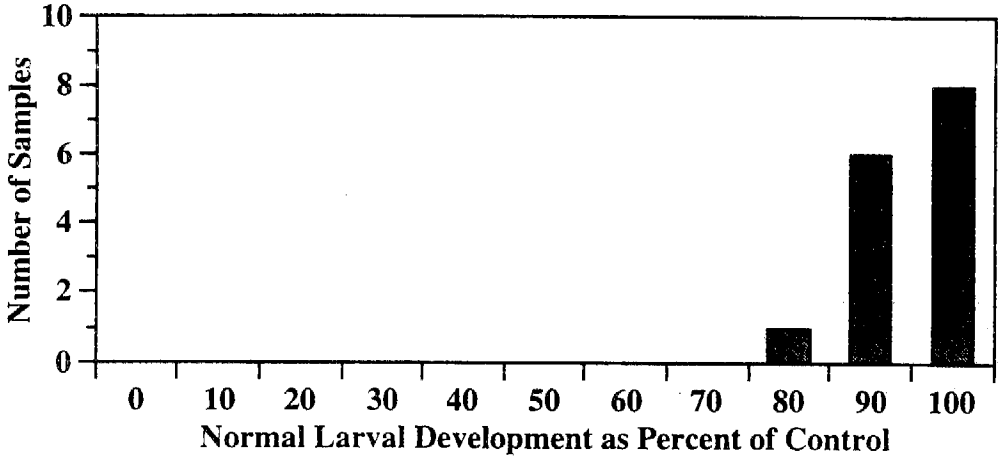


Figure 36b. Reference envelope tolerance limits for the sea urchin test at the sediment/water interface. Data are from this study plus additional BPTCP and RMP studies.There were no outliers identified or removed from this data set.

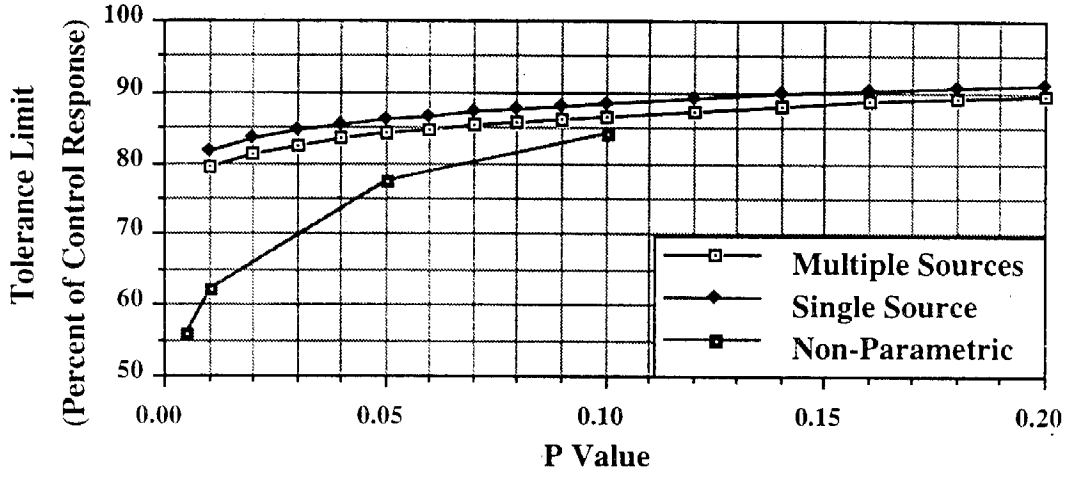


Figure 37a. Distribution of reference site data for the *Neanthes* test in homogenized sediment. All data were from this study. There were no outliers identified or removed from this data set.

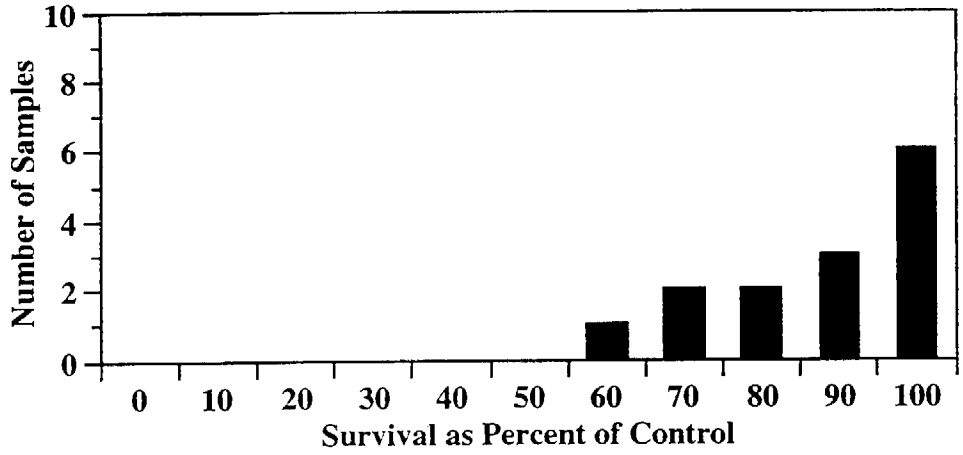


Figure 37b. Reference envelope tolerance limits for the *Neanthes* survival test. All data were from this study. There were no outliers identified or removed from this data set.

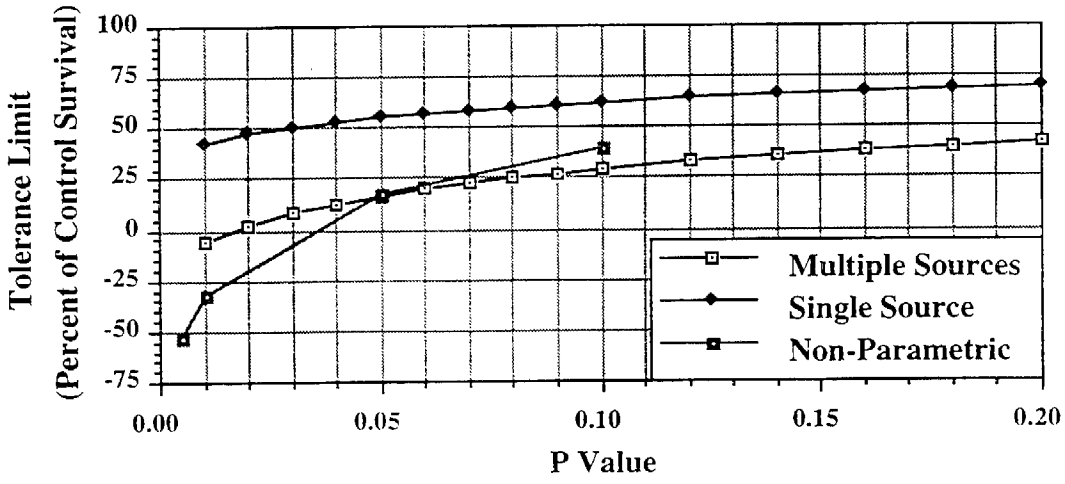


Figure 38a. Distribution of reference site data for the *Neanthes* growth test in homogenized sediment. All data were from this study. There were no outliers identified or removed from this data set.

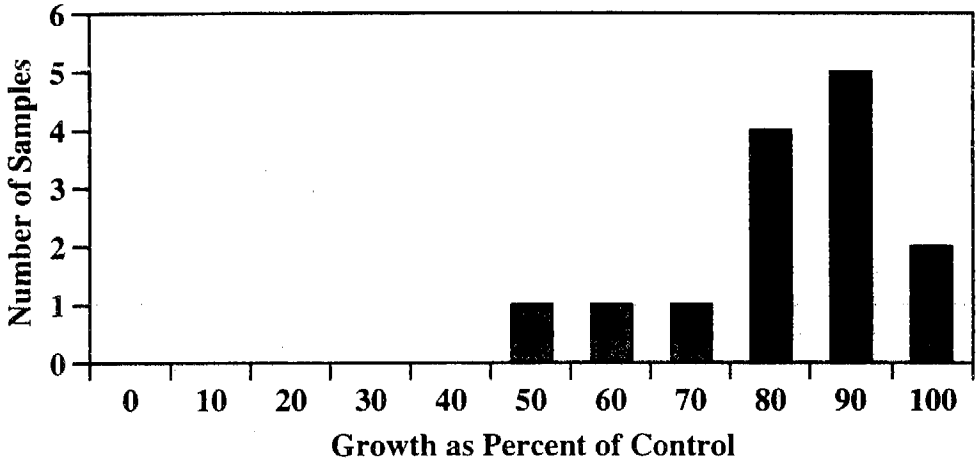


Figure 38b. Reference envelope tolerance limits for the *Neanthes* growth test. All data were from this study. There were no outliers identified or removed from this data set.

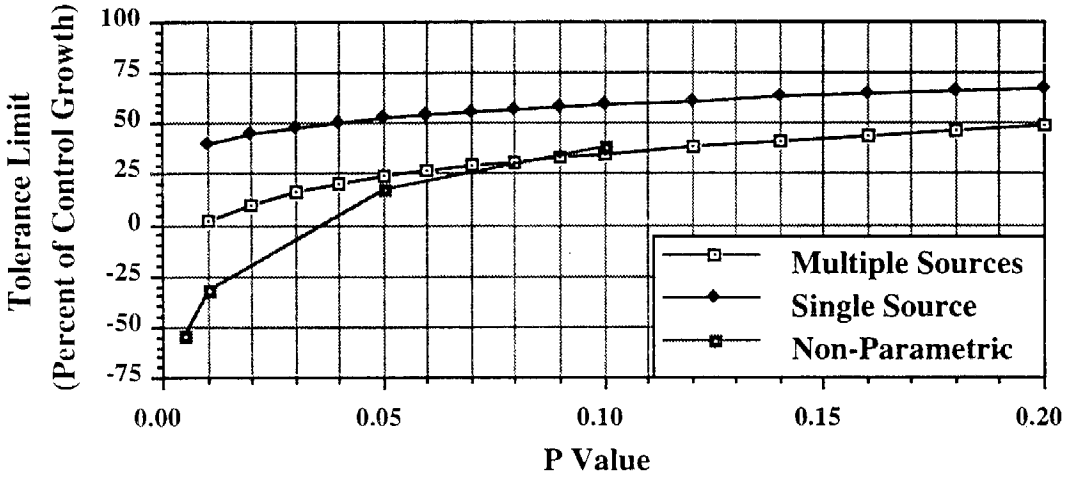


Figure 39a. Distribution of reference site data for the *Nebalia* test in homogenized sediment. All data were from this study. There were no outliers identified or removed from this data set.

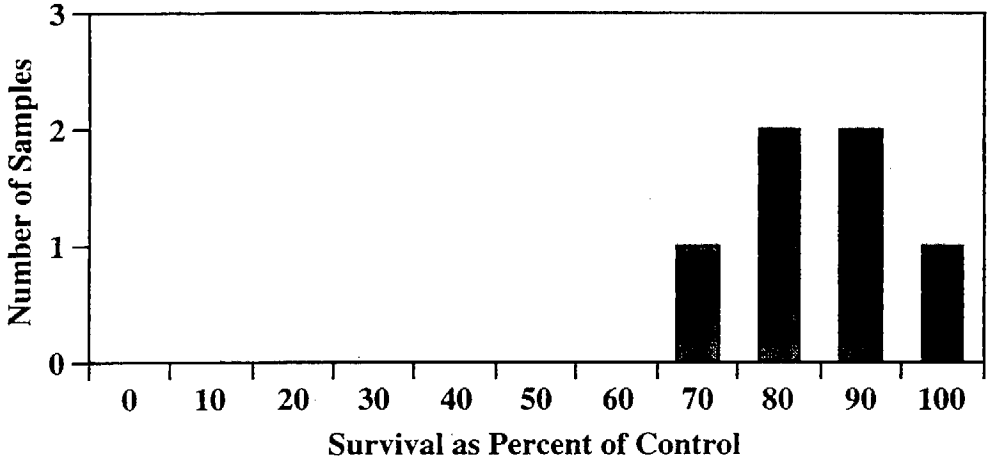


Figure 39b. Reference envelope tolerance limits for the *Nebalia* survival test. All data were from this study. There were no outliers identified or removed from this data set.

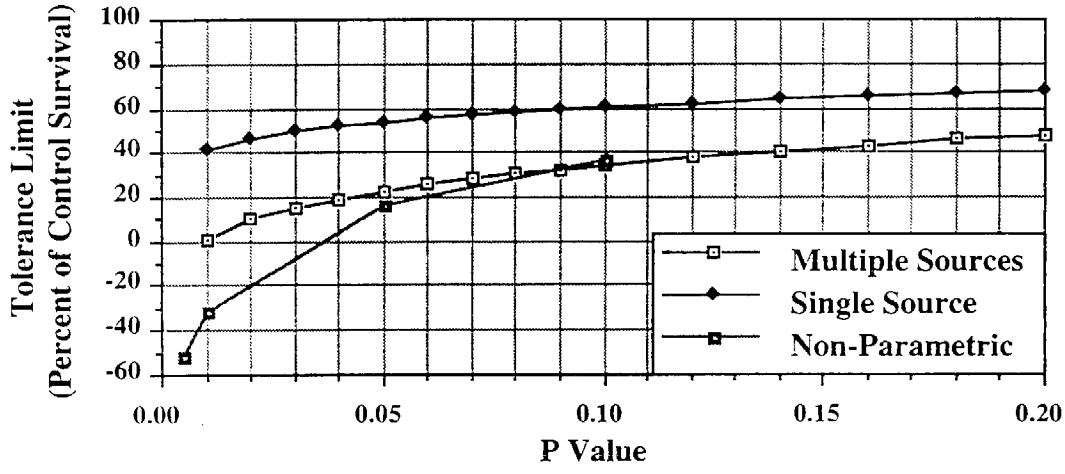


Table 9a. Summary of trace metal chemistry data for samples with concentrations exceeding PEL and/or ERM values.

Station	Date	Mean ERMQ	Mean PELQ	Cr PELQ	Cr Conc.	Ni ERMQ	Ni PELQ	Ni Conc.	Mean Metal ERMQ	Mean Metal PELQ
Paradise Cove (1)	4/94	0.22	0.35	1.5	238	2.1	2.5	107	0.30	0.48
Paradise Cove (2)	4/94	0.19	0.32	1.4	219	1.8	2.2	93	0.25	0.42
Paradise Cove (3)	4/94	0.19	0.31	1.4	222	2.0	2.4	104	0.26	0.42
San Pablo Bay, Tubbs Is. (1)	4/94	0.22	0.33	1.3	207	2.4	2.9	123	0.31	0.47
San Pablo Bay, Tubbs Is. (2)	4/94	0.20	0.31	1.2	195	2.2	2.6	113	0.28	0.42
San Pablo Bay, Tubbs Is. (3)	4/94	0.19	0.30	1.2	198	2.3	2.7	117	0.27	0.42
San Pablo Bay, Is. #1 (1)	4/94	0.18	0.29	1.2	194	1.4	1.7	73	0.25	0.41
San Pablo Bay, Is. #1 (2)	4/94	0.20	0.30	1.3	202	1.5	1.8	76	0.28	0.43
San Pablo Bay, Is. #1 (3)	4/94	0.20	0.31	1.2	195	1.4	1.7	73	0.28	0.43
Paradise Cove (1)	3/95	0.21	0.35	1.2	196	1.9	2.3	98	0.24	0.40
N. South Bay (1)	3/95	0.17	0.28	1.1	181	2.0	2.4	102	0.22	0.37
N. South Bay (2)	3/95	0.17	0.28	1.2	186	1.9	2.3	98	0.22	0.36
N. South Bay (3)	3/95	0.17	0.28	1.2	193	1.9	2.3	96	0.22	0.37
S. South Bay (1)	3/95	0.16	0.29	1.3	212	1.7	2.0	85	0.20	0.37
N. South Bay (2)	3/95	0.16	0.29	1.3	213	1.6	1.9	83	0.20	0.36
N. South Bay (3)	3/95	0.16	0.29	1.3	206	1.6	1.9	83	0.20	0.36
San Pablo Bay, Tubbs Is. (1)	4/95	0.21	0.34	1.3	209	2.6	3.2	135	0.29	0.48
San Pablo Bay, Is. #1 (1)	4/95	0.27	0.37	1.1	181	2.0	2.4	102	0.24	0.39

Trace metal concentration units are ppm (µg/g dry weight).
ERM is Effects Range Median, PEL is Probable Effects Level (see section 10.2.1).
ERMQ and PELQ are quotients: (measured concentration of a chemical) ÷ (its ERM or PEL value).
Mean ERMQ and PELQ are averages of ERM quotient or PEL quotient values for all measured chemicals (metal & organic), except nickel (Ni).
Mean Metal ERMQ and PELQ are average quotient values for all measured trace metals, except nickel (Ni; see Results Section 3.0).
Cr is measured chromium, Cr PELQ is the measured chromium concentration divided by the PEL value (the ERM was not exceeded).
Ni is measured nickel, Ni ERMQ and Ni PELQ are the measured nickel concentrations divided by the ERM and PEL values.

Table 9b. Summary of trace organic chemistry data for samples with concentrations exceeding PEL and/or ERM values.

Station	Date	Mean ERMQ	Mean PELQ	ppDDT PELQ	ppDDT Conc.	Total DDT ERMQ	Total DDT PELQ	Total DDT Conc.	DBA ERMQ	DBA PELQ	DBA Conc.
Paradise Cove (1)	4/94	0.22	0.35	0.1	nd	0.2	0.1	6.96	0.1	0.2	30.2
Paradise Cove (2)	4/94	0.19	0.32	0.1	nd	0.1	0.1	6.47	0.1	0.2	26.0
Paradise Cove (3)	4/94	0.19	0.31	0.1	nd	0.1	0.1	5.78	0.1	0.2	24.9
San Pablo Bay, Tubbs Is. (1)	4/94	0.22	0.33	0.1	nd	0.1	0.1	5.74	0.1	0.1	18.0
San Pablo Bay, Tubbs Is. (2)	4/94	0.20	0.31	0.1	nd	0.1	0.1	6.24	0.1	0.1	18.5
San Pablo Bay, Tubbs Is. (3)	4/94	0.19	0.30	0.1	nd	0.1	0.1	6.18	0.1	0.1	17.9
San Pablo Bay, Is. #1 (1)	4/94	0.18	0.29	0.1	nd	0.1	0.1	5.76	0.1	0.1	13.3
San Pablo Bay, Is. #1 (2)	4/94	0.20	0.30	0.1	nd	0.1	0.1	6.07	0.1	0.1	13.4
San Pablo Bay, Is. #1 (3)	4/94	0.20	0.31	0.1	nd	0.1	0.1	5.95	0.1	0.1	18.0
Paradise Cove (1)	3/95	0.21	0.35	0.1	nd	0.1	0.1	6.50	0.5	1.010	136.0
N. South Bay (1)	3/95	0.17	0.28	0.1	nd	0.1	0.1	4.61	0.1	0.2	23.3
N. South Bay (2)	3/95	0.17	0.28	0.1	nd	0.1	0.1	4.39	0.2	0.3	44.5
N. South Bay (3)	3/95	0.17	0.28	0.1	nd	0.1	0.1	4.43	0.1	0.2	30.5
S. South Bay (1)	3/95	0.16	0.29	0.1	nd	0.1	0.1	4.76	0.1	0.2	30.0
N. South Bay (2)	3/95	0.16	0.29	0.1	nd	0.1	0.1	3.71	0.1	0.3	34.6
N. South Bay (3)	3/95	0.16	0.29	0.1	nd	0.1	0.1	3.34	0.1	0.3	34.1
San Pablo Bay, Tubbs Is. (1)	4/95	0.21	0.34	0.1	nd	0.1	0.1	6.61	0.1	0.2	23.1
San Pablo Bay, Is. #1 (1) *	4/95	0.27	0.37	12.1*	58.1*	1.6*	1.4*	71.78*	0.1	0.1	19.9

Trace organic concentration units are ppb (ng/g dry weight). ERM is Effects Range Median, PEL is Probable Effects Level (see section 10.2.1).
ERMQ and PELQ are quotients: (measured concentration of a chemical) ÷ (its ERM or PEL value).
Mean ERMQ and PELQ are averages of ERM quotient or PEL quotient values for all measured chemicals, except nickel (Ni, see Results Section 3.0).
DBA is Dibenz[a,h]anthracene. ppDDT is p',p' DDT, for which there are no ERM guidelines. *See Results Section 2.0 regarding these DDT values.*
Total DDT is the sum of [o'p'DDD], [p'p'DDD], [o'p' DDE], [p'p' DDE], [o'p' DDT] and [p'p' DDT];
*ERM & PEL quotient sums calculated using [total DDT], not DDT metabolite quotients. nd = non-detected.

Table 10. Toxicity data summary. Toxicity test results (mean ± sd) for each protocol and endpoint used in this study. Data are not presented as a percent of control in this table, as they are elsewhere in the Results section. Sample results corresponded to controls marked "(1)", with the following exceptions: fine sediment controls were used if available for Ampelisca tests; controls marked "(2)" were used for the two San Pablo Bay sites in the 3/95 tests; brine controls were used for pore water samples in which salinity was adjusted; and brine controls "(2)" were used for San Pablo Bay pore water in 4/94 tests.

Site Name	Site Number	Field Rep.	Ampelisca Homogenate			Eohaustorius Homogenate		
			% surv 4/94	% surv 9/94	% surv 3/95	% surv 4/94	% surv 9/94	% surv 3/95
San Pablo Bay	20007	1	85 ± 17	74 ± 8	89 ± 7	90 ± 12	88 ± 12	85 ± 10
(Island #1)		2	82 ± 10	74 ± 4	94 ± 11	85 ± 4	86 ± 12	90 ± 5
		3	92 ± 9	61 ± 8	92 ± 8	95 ± 7	80 ± 6	77 ±12
San Pablo Bay	20006	1	94 ± 11	55 ± 18	79 ± 10	72 ± 8	66 ± 39	80 ± 4
(Tubbs Isl.)		2	95 ± 7	79 ± 6	81 ± 10	70 ± 6	62 ± 35	80 ± 10
		3	89 ± 10	69 ± 28	82 ± 10	78 ± 6	72 ± 18	81 ± 6
Paradise Cove	20005	1	82 ± 17	69 ± 10	97 ± 5	79 ± 13	82 ± 6	82 ± 11
		2	85 ± 9	76 ± 12	97 ± 4	75 ± 6	81 ± 16	86 ± 13
		3	80 ± 20	89 ± 6	94 ± 7	79 ± 11	84 ± 13	85 ± 8
N. South Bay	20013	1			83 ± 14			76 ± 11
		2			98 ± 3			82 ± 14
		3			95 ± 5			70 ± 17
S. South Bay	20014	1			87 ± 14			57 ± 34
		2			86 ± 9			89 ± 4
		3			84 ± 11			68 ± 39
Tomales Bay	20009	1	73 ± 10	79 ± 7		32 ± 31	78 ± 10	
(Marconi Cove)		2	76 ± 11	82 ± 6	91 ± 11	53 ± 19	54 ± 32	67 ± 8
		3	73 ± 10	78 ± 10		65 ± 7	85 ± 11	
Bolinas Lagoon	20008	1	82 ± 8			83 ± 10		
(Audubon Cyn)		2	77 ± 26			90 ± 11		
		3	80 ± 15			75 ± 22		
Castro Cove	20010	1		29 ± 14			33 ± 3	
Clipper Cove	20012	1			90 ± 7			80 ± 15
Islais Creek	20011	1		54 ± 19			57 ± 14	
Controls								
Home (1)			80 ± 13		91 ± 9	93 ± 8	95 ± 4	92 ± 7
Home (2)					96 ± 6			97 ± 7
Home (fine sed)			85 ± 12	81 ± 10				
Dilution (1)								
Brine (1)								
Dilution (2)								
Brine (2)								

Table 10 (Continued). Toxicity data summary. Toxicity test results (mean ± sd) for each protocol used in this study. Data are not presented as a percent of control in this table, as they are elsewhere in the Results section. Sample results corresponded to controls marked "(1)", with the following exceptions: fine sediment controls were used if available for Ampelisca tests; controls marked "(2)" were used for the two San Pablo Bay sites in the 3/95 tests; brine controls were used for pore water samples in which salinity was adjusted; and brine controls "(2)" were used for San Pablo Bay pore water in 4/94 tests.

Site Name	Site Number	Field Rep.	Eohaustorius Intact		Eohaustorius Pore Water		
			% surv 4/94	% surv 9/94	% surv 4/94	% surv 9/94	% surv 3/95
San Pablo Bay (Island #1)	20007	1	90 ± 9	72 ± 16	68 ± 23	84 ± 9	92 ± 11
		2			72 ± 11	96 ± 9	96 ± 9
		3			76 ± 22	88 ± 18	96 ± 9
San Pablo Bay (Tubbs Isl.)	20006	1	70 ± 12	19 ± 37	76 ± 26	88 ± 18	80 ± 20
		2			68 ± 23	84 ± 17	92 ± 11
		3			68 ± 23	88 ± 18	96 ± 9
Paradise Cove	20005	1	63 ± 29	60 ± 31	76 ± 9	96 ± 9	84 ± 26
		2			56 ± 30	76 ± 17	76 ± 17
		3			84 ± 17	84 ± 9	60 ± 32
N. South Bay	20013	1					64 ± 17
		2					52 ± 36
		3					84 ± 26
S. South Bay	20014	1					72 ± 18
		2					72 ± 11
		3					64 ± 9
Tomales Bay (Marconi Cove)	20009	1	53 ± 29	82 ± 13	68 ± 23	96 ± 9	
		2			68 ± 18	72 ± 11	52 ± 27
		3			52 ± 23	88 ± 11	
Bolinas Lagoon (Audubon Cyn)	20008	1	45 ± 25		80 ± 20		
		2			76 ± 9		
		3			76 ± 26		
Castro Cove	20010	1		34 ± 22		44 ± 33	
Clipper Cove	20012	1					48 ± 23
Islais Creek	20011	1		41 ± 27		0 ± 0	
Controls							
Home (1)			97 ± 5	95 ± 4			
Home (2)							
Home (fine sed)							
Dilution (1)					80 ± 14	84 ± 17	48 ± 23
Brine (1)							48 ± 30
Dilution (2)							84 ± 9
Brine (2)							84 ± 17

Table 10 (Continued). Toxicity data summary. Toxicity test results (mean ± sd) for each protocol used in this study. Data are <u>not</u> presented as a percent of control in this table, as they are elsewhere in the Results section. Sample results corresponded to controls marked "(1)", with the following exceptions: fine sediment controls were used if available for Ampelisca tests; controls marked "(2)" were used for the two San Pablo Bay sites in the 3/95 tests; brine controls were used for pore water samples in which salinity was adjusted; and brine controls "(2)" were used for San Pablo Bay pore water in 4/94 tests.

Site Name	Site Number	Field Rep.	Mussel Pore Water			Sea Urchin Pore Water		
			% normal 4/94	% normal 9/94	% normal 3/95	% normal 4/94	% normal 9/94	% normal 3/95
San Pablo Bay (Island #1)	20007	1	79 ± 11	41 ± 7	78 ± 9	98 ± 1	97 ± 2	95 ± 2
		2	83 ± 9	52 ± 7	76 ± 7	96 ± 3	94 ± 3	98 ± 2
		3	81 ± 11	45 ± 6	78 ±8	94 ± 3	97 ± 2	96 ± 1
San Pablo Bay (Tubbs Isl.)	20006	1	83 ± 9	22 ± 7	78 ± 8	96 ± 2	96 ± 3	98 ± 1
		2	77 ± 13	34 ± 7	76 ± 6	97 ± 1	97 ± 2	96 ± 3
		3	93 ± 2	31 ± 8	78 ±6	95 ± 2	97 ± 1	97 ± 1
Paradise Cove	20005	1	69 ± 20	33 ± 6	93 ± 3	92 ± 4	94 ± 5	95 ± 2
		2	65 ± 12	29 ± 11	89 ± 3	90 ± 4	96 ± 2	97 ± 1
		3	77 ± 5	37 ± 6	78 ± 7	93 ± 6	96 ± 3	98 ± 1
N. South Bay	20013	1			89 ± 5			94 ± 3
		2			90 ± 7			98 ± 1
		3			82 ± 20			95 ± 4
S. South Bay	20014	1			90 ± 6			98 ± 0
		2			92 ± 5			99 ± 1
		3			87 ± 4			97 ± 1
Tomales Bay (Marconi Cove)	20009	1	72 ± 9	25 ± 5		97 ± 2	82 ± 4	
		2	78 ± 12	34 ± 6	85 ± 6	92 ± 7	64 ± 21	91 ± 7
		3	83 ± 12	24 ± 10		72 ± 41	71 ± 7	
Bolinas Lagoon (Audubon Cyn)	20008	1	85 ± 10			95 ± 3		
		2	79 ± 11			95 ± 4		
		3	86 ± 10			95 ± 2		
Castro Cove	20010	1		28 ± 5			96 ± 1	
Clipper Cove	20012	1			92 ± 5			94 ± 3
Islais Creek	20011	1		0 ± 0			0 ± 0	
Controls								
Home (1)								
Home (2)								
Home (fine sed)								
Dilution (1)			76 ± 8	59 ± 4	94 ± 7	95 ± 4	96 ± 2	98 ± 2
Brine (1)			73 ± 7	60 ± 7	90 ± 7	90 ± 4	92 ± 4	96 ± 2
Dilution (2)			80 ± 8		84 ± 10	95 ± 2		99 ± 2
Brine (2)			70±17		83 ± 7	94 ± 2		98 ± 2

Table 10 (Continued). Toxicity data summary. Toxicity test results (mean ± sd) for each protocol used in this study. Data are not presented as a percent of control in this table, as they are elsewhere in the Results section. Sample results corresponded to controls marked "(1)", with the following exceptions: fine sediment controls were used if available for Ampelisca tests; controls marked "(2)" were used for the two San Pablo Bay sites in the 3/95 tests; brine controls were used for pore water samples in which salinity was adjusted; and brine controls "(2)" were used for San Pablo Bay pore water in 4/94 tests.

Site Name	Site Number	Field Rep.	Sea Urchin SWI			Neanthes Homogenate		
			% normal 4/94	% normal 9/94	% normal 3/95	% surv 9/94	gwth (mg) 9/94	gwth (%) 9/94
San Pablo Bay (Island #1)	20007	1	93 ± 2	98 ± 2	84 ± 7	100 ± 0	11.2 ± 1.2	88.2
		2				100 ± 0	12.4 ± 2.8	97.6
		3				76 ± 43	13.7 ± 4.4	107.9
San Pablo Bay (Tubbs Isl.)	20006	1	95 ± 2	94 ± 2	93 ± 4	76 ± 43	11.5 ± 1.4	90.6
		2				80 ± 28	11.8 ± 3.4	92.9
		3				96 ± 9	12.2 ± 2.4	96.1
Paradise Cove	20005	1	92 ± 3	95 ± 2	96 ± 1	88 ± 27	11.1 ± 3.6	87.4
		2				60 ± 42	13.1 ± 4.0	103.1
		3				92 ± 18	9.9 ± 3.6	78
N. South Bay	20013	1			97 ± 1			
		2						
		3						
S. South Bay	20014	1			98 ± 1			
		2						
		3						
Tomales Bay (Marconi Cove)	20009	1	84 ± 25	77 ± 43		56 ± 30	12.2 ± 5.1	96.1
		2			96 ± 2	96 ± 9	10.3 ± 3.1	81.1
		3				80 ± 45	10.2 ± 3.0	80.3
Bolinas Lagoon (Audubon Cyn)	20008	1	95 ± 4					
		2						
		3						
Castro Cove	20010	1		97 ± 3		100 ± 0	12.5 ± 3.5	98.4
Clipper Cove	20012	1			95 ± 5			
Islais Creek	20011	1		0 ± 0		100 ± 0	11.3 ± 4.1	89
Controls								
Home (1)			99 ± 1			100 ± 0	12.7 ± 2.5	100
Home (2)								
Home (fine sed)								
Dilution (1)				93 ± 2	98 ± 1			
Brine (1)					97 ± 1			
Dilution (2)								
Brine (2)					93 ± 1			

Table 10 (Continued). Toxicity data summary. Toxicity test results (mean ± sd) for each protocol used in this study. Data are not presented as a percent of control in this table, as they are elsewhere in the Results section. Sample results corresponded to controls marked "(1)", with the following exceptions: fine sediment controls were used if available for Ampelisca tests; controls marked "(2)" were used for the two San Pablo Bay sites in the 3/95 tests; brine controls were used for pore water samples in which salinity was adjusted; and brine controls "(2)" were used for San Pablo Bay pore water in 4/94 tests.

Site Name	Site Number	Field Rep.	Neanthes Homogenate			Nebalia Homog.	
			% surv 3/95	gwth (mg) 3/95	gwth (%) 3/95	% surv 9/94	% surv 3/95
San Pablo Bay (Island #1)	20007	1 2 3	100 ± 0	12.6 ± 2.8	59.0	95 ± 6	
San Pablo Bay (Tubbs Isl.)	20006	1 2 3	100 ± 0	14.1 ± 0.5	66.0	93 ± 11	
Paradise Cove	20005	1 2 3	80 ± 45	13.4 ± 2.9	85.0	100 ± 0	65 ± 15
N. South Bay	20013	1 2 3	96 ± 9	12.8 ± 3.2	82.0		69 ± 10
S. South Bay	20014	1 2 3	92 ± 11	15.3 ± 4.5	97.0		70 ± 16
Tomales Bay (Marconi Cove)	20009	1 2 3	100 ± 0	10.4 ± 2.3	66.0	97 ± 3	81 ± 6
Bolinas Lagoon (Audubon Cyn)	20008	1 2 3					
Castro Cove	20010	1				100 ± 0	
Clipper Cove	20012	1	100 ± 0	10.5 ± 1.9	67.0		72 ± 17
Islais Creek	20011	1				97 ± 5	
Controls							
Home (1)			88 ± 11	15.7 ± 1.4	100.0	100 ± 0	85 ± 4
Home (2)			100 ± 0	21.4 ± 2.1	100.0		
Home (fine sed)							
Dilution (1)							
Brine (1)							
Dilution (2)							
Brine (2)							

Table 11. Spearman rank correlation coefficients for significant negative correlations between toxicity and natural sediment parameters.

Test	TOC	% Clay	% Fines	NH3	H2S
Ampelisca	NS	NS	NS	NS	NS
Eohaustorius (Homog)	-0.570 ***	-0.321 *	NS	NS	NS
Eohaustorius (Intact)	NS	NS	NS	NS	NS
Eohaustorius (Pore Water)	-0.347 **	na	na	-0.716 ***	-0.681 ***
Mytilus	NS	na	na	NS	NS
Sea Urchin (Pore Water)	-0.333 *	na	na	NS	NS
Sea Urchin (SWI)	NS	na	na	NS	NS
Neanthes (Survival)	NS	NS	-0.475 *	NS	NS
Neanthes (Growth)	NS	NS	NS	NS	NS
Nebalia	NS	NS	NS	NS	NS

NS = not significant. na = not applicable (e.g. grain size in pore water tests).
TOC is total organic carbon.
Statistical significance: alpha 0.05*; alpha 0.01**; alpha 0.001***

Table 12. Toxicity test sulfide and ammonia measurements above threshold values (a).

Site	Site Number	Date	Test	Measurements from Test Chambers			
				Parameter (b)	Time (c)	Matrix (d)	Concentration (mg/L)
SPB Island #1	20007	Apr-94	Mussel PW	S2-	I	I	0.09
SPB Island #1	20007	Apr-94	Urchin PW	S2-	I	I	0.14
SPB Island #1	20007	Apr-94	Urchin PW	S2-	I	I	0.159
Bolinas Lagoon	20008	Apr-94	Mussel PW	S2-	I	I	0.106
Tomales Bay	20009	Apr-94	Mussel PW	S2-	I	I	0.166
Tomales Bay	20009	Apr-94	Mussel PW	S2-	I	I	0.113
Tomales Bay	20009	Apr-94	Urchin PW	S2-	I	I	0.17
Islais Creek	20011	Sep-94	Ampelisca	NH3	F	O	0.721
Islais Creek	20011	Sep-94	Ampelisca	S2-	I	I	3.967
Islais Creek	20011	Sep-94	Eoh Homog	S2-	I	I	6.164
Islais Creek	20011	Sep-94	Eoh Intact	S2-	I	I	4.956
Islais Creek	20011	Sep-94	Eoh Intact	S2-	F	I	2.349
Islais Creek	20011	Sep-94	Eoh PW	S2-	I	I	1.373
Islais Creek	20011	Sep-94	Urchin PW	NH3	F	I	0.478
Islais Creek	20011	Sep-94	Urchin PW	S2-	F	I	0.935
Islais Creek	20011	Sep-94	Urchin SWI	NH3	F	O	0.083
N South Bay	20013	Mar/Apr 95	Mussel PW	NH3	I	I	0.057
N South Bay	20013	Mar/Apr 95	Nebalia	NH3	I	O	1.835
N South Bay	20013	Mar/Apr 95	Urchin SWI	NH3	I	I	0.079
N South Bay	20013	Mar/Apr 95	Urchin SWI	NH3	F	I	0.054

(a) Threshold values for Total Sulfide were derived from the following sources:

Ampelisca LOEC (for Rhepoxynius) = 1.47 mg/L, Knezovich et al 1995.
Eohaustorius LOEC for Eohaustorius = 1.92 mg/L, Knezovich et al 1995.
Neanthes LOEC for Neanthes = 10 mg/L, Dillon et al 1993
Nebalia LOEC (for Rhepoxynius) = 1.47 mg/L, Knezovich et al 1995.
Sea Urchin LOEC for S. purpuratus = 0.128 mg/L, Knezovich et al 1995.
Mussel LOEC for M. edulis = 0.09, Knezovich et al 1995.

(a) Threshold values for Unionized Ammonia derived from the following sources:

Ampelisca Toxicity test application limit = 0.4 mg/L, EPA 1994.
Eohaustorius Toxicity test application limit = 0.8 mg/L, EPA 1994.
Neanthes LOEC for Neanthes = 1.25 mg/L, Dillon et al 1993
Nebalia Toxicity test application limit (for Rhepoxynius) = 0.4 mg/L, EPA 1994.
Sea Urchin NOEC for S. purpuratus = 0.05 mg/L, Bay et al 1993.
Mussel NOEC (for red abalone larvae) = 0.05 mg/L, MPSL, unpublished data.

(b) S2- is total sulfide, NH3 is unionized ammonia.
(c) "I" indicates measurement taken at test initiation, "F" is final at test termination.
(d) "I" indicates measurement taken from interstitial water,"O" is from overlying water.

Table 13. Spearman rank correlation coefficients for significant negative correlations between toxicity and bulk sediment chemistry.

Test	As	Cu	Fe	Sb	Zn	ppDDE	Total PCB	TOC
Ampelisca	NS	NS	NS	NS	NS	NS	NS	NS
Eohaustorius H.	-0.570 ***	-0.321 *	NS	NS	NS	NS	NS	-0.570 ***
Eohaustorius I.	NS	NS	NS	NS	NS	NS	NS	NS
Eohaustorius PW	NS	-0.672 ***	-0.535 ***	-0.618 ***	-0.581 ***	-0.607 ***	NS	-0.347 **
Mytilus PW	NS	NS	NS	NS	NS	NS	NS	NS
Sea Urchin PW	NS	NS	NS	NS	NS	NS	-0.608 ***	-0.333 *
Sea Urchin SWI	-0.880	NS	NS	NS	NS	-0.755 ***	NS	NS
Neanthes Surv.	NS	NS	NS	NS	NS	NS	NS	NS
Neanthes Grow.	NS	NS	NS	NS	NS	NS	NS	NS
Nebalia	NS	NS	NS	NS	NS	NS	NS	NS

NS indicates the correlation was not statistically significant.

As is arsenic, Cu is copper, Fe is iron, Sb is antimony, Zn is zinc, ppDDE is p',p' DDE, and total PCB is the sum of 18 PCB congeners.

TOC is total organic carbon.

Statistical significance: alpha 0.05*; alpha 0.01**; alpha 0.001***

Table 14. Data included in additional tolerance limit calculations. All data are from candidate reference sites in San Francisco Bay. BPTCP is Bay Protection and Toxic Cleanup Program; these reference sites are the same as those sampled as part of this study, and were sampled in conjunction with toxicity screening of Bay test sites. RMP is the SF Bay Regional Monitoring Program; these sites were sampled as part of semi-annual Bay surveys. SWI indicates sediment water interface exposures (see Methods Section 3.6.2).

Site Name	Site Code	Date Collected	Percent of Control Response		
			Eohaustorius Survival	Sea Urchin Normal Larvae	
				Pore Water	SWI
BPTCP Sites					
N. South Bay	20013	4/19/95	91%	122%	
S. South Bay	20014	4/19/95	88%	55%	
Island # 1	20007	5/2/95	85%	101%	
Paradise Cove	20005	5/1/95	85%	100%	
Tubbs Island	20006	10/26/95	91%	105%	
Paradise Cove	20005	10/26/95	86%	3%	
N. South Bay	20013	12/7/95	87%	0%	
S. South Bay	20014	12/7/95	89%	102%	
Tubbs Island	20006	6/11/96			103%
Paradise Cove	20005	4/4/97	79%		97%
N. South Bay	20013	4/16/97	100%		100%
Island # 1	20007	4/15/97	52%		90%
RMP Sites					
Pinole Point	BD30	3/1/93	64%		
Pinole Point	BD30	9/1/93	89%		
Pinole Point	BD30	2/1/94	74%		
Horseshoe Bay	BC21	2/1/94	86%		
San Bruno Shoal	BB15	8/1/94	100%		
Horseshoe Bay	BC21	8/1/94	101%		
San Bruno Shoal	BB15	2/1/95	80%		
Horseshoe Bay	BC21	2/1/95	90%		
San Bruno Shoal	BB15	8/1/95	83%		
Horseshoe Bay	BC21	8/1/95	89%		
San Bruno Shoal	BB15	2/1/96	84%		
Horseshoe Bay	BC21	2/1/96	76%		
San Bruno Shoal	BB15	8/1/96	90%		
Horseshoe Bay	BC21	8/1/96	88%		
San Bruno Shoal	BB15	2/1/97	83%		
Horseshoe Bay	BC21	2/1/97	82%		

Table 15. Tolerance limits, presented as survival or normal development as a percentage of test controls, based on reference site toxicity data from this study, BPTCP screening studies, and the RMP, with outliers removed. The "p" value indicates the percentile of the reference distribution used to generate the tolerance limit. Tolerance limits based on calculations using multiple sources of variation are appropriate for the current study; non-parametric limits and limits based on calculations using a single source of variation are shown for comparison. All limits were calculated based on an alpha level of 0.05. See Methods Section 10.3 for details. "nc" indicates limit was not calculated.

Test	p value	Tolerance Limits		
		Parametric		Non-Parametric
		Sources of Variation:		
		Multiple	Single	
Ampelisca	1%	54.7	64.7	28.7
Ampelisca	2%	59.1	68.1	nc
Ampelisca	3%	61.6	70.3	nc
Ampelisca	4%	63.7	72.0	nc
Ampelisca	5%	65.3	73.3	57.9
Ampelisca	6%	66.6	74.4	nc
Ampelisca	7%	67.9	75.4	nc
Ampelisca	8%	68.9	76.3	nc
Ampelisca	9%	69.9	77.1	nc
Ampelisca	10%	70.9	77.8	75.3
Ampelisca	12%	72.5	79.1	nc
Ampelisca	14%	73.9	80.2	nc
Ampelisca	16%	75.1	81.3	nc
Ampelisca	18%	76.3	82.2	nc
Ampelisca	20%	77.5	83.1	nc
Eohaustorius Homog.	1%	58.7	61.0	40.4
Eohaustorius Homog.	2%	61.5	63.7	nc
Eohaustorius Homog.	3%	63.3	65.4	nc
Eohaustorius Homog.	4%	64.2	66.7	nc
Eohaustorius Homog.	5%	65.5	67.7	63.9
Eohaustorius Homog.	6%	66.7	68.6	nc
Eohaustorius Homog.	7%	67.5	69.3	nc
Eohaustorius Homog.	8%	68.2	70.0	nc
Eohaustorius Homog.	9%	68.8	70.7	nc
Eohaustorius Homog.	10%	69.5	71.2	65.3
Eohaustorius Homog.	12%	70.6	72.3	nc
Eohaustorius Homog.	14%	71.5	73.2	nc
Eohaustorius Homog.	16%	72.2	74.0	nc

Table 15. Continued.

Test	p value	Tolerance Limits		
		Parametric		Non-Parametric
		Sources of Variation:		
		Multiple	Single	
Eohaustorius Homog.	18%	72.8	74.7	nc
Eohaustorius Homog.	20%	73.4	75.4	nc
Eohaustorius Intact	1%	-2.1	5.9	-111.4
Eohaustorius Intact	2%	5.4	13.3	nc
Eohaustorius Intact	3%	10.9	17.9	nc
Eohaustorius Intact	4%	14.7	21.3	nc
Eohaustorius Intact	5%	17.3	24.1	-37.1
Eohaustorius Intact	6%	19.7	26.5	nc
Eohaustorius Intact	7%	22.5	28.5	nc
Eohaustorius Intact	8%	24.5	30.3	nc
Eohaustorius Intact	9%	26.3	32.0	nc
Eohaustorius Intact	10%	27.6	33.5	-5.2
Eohaustorius Intact	12%	30.6	36.2	nc
Eohaustorius Intact	14%	33.1	38.6	nc
Eohaustorius Intact	16%	35.5	40.7	nc
Eohaustorius Intact	18%	37.6	42.6	nc
Eohaustorius Intact	20%	39.5	44.4	nc
Eohaustorius Pore Water	1%	-39.1	46.2	-13.3
Eohaustorius Pore Water	2%	-25.1	53.1	nc
Eohaustorius Pore Water	3%	-14.9	57.4	nc
Eohaustorius Pore Water	4%	-7.9	60.7	nc
Eohaustorius Pore Water	5%	-2.7	63.3	47.1
Eohaustorius Pore Water	6%	0.8	65.5	nc
Eohaustorius Pore Water	7%	5.0	67.5	nc
Eohaustorius Pore Water	8%	11.0	69.2	nc
Eohaustorius Pore Water	9%	14.4	70.8	nc
Eohaustorius Pore Water	10%	17.0	72.3	85.0
Eohaustorius Pore Water	12%	23.5	74.9	nc
Eohaustorius Pore Water	14%	29.3	77.2	nc
Eohaustorius Pore Water	16%	32.8	79.2	nc
Eohaustorius Pore Water	18%	36.6	81.1	nc
Eohaustorius Pore Water	20%	40.2	82.9	nc
Mussel Larvae	1%	19.9	65.2	32.6
Mussel Larvae	2%	28.4	69.2	nc
Mussel Larvae	3%	34.1	71.7	nc
Mussel Larvae	4%	37.1	73.6	nc

Table 15. Continued.

Test	p value	Tolerance Limits		
		Parametric		Non-Parametric
		Sources of Variation:		
		Multiple	Single	
Mussel Larvae	5%	39.8	75.1	67.9
Mussel Larvae	6%	42.2	76.4	nc
Mussel Larvae	7%	44.8	77.6	nc
Mussel Larvae	8%	45.6	78.6	nc
Mussel Larvae	9%	48.4	79.5	nc
Mussel Larvae	10%	50.7	80.3	83.2
Mussel Larvae	12%	54.0	81.8	nc
Mussel Larvae	14%	56.4	83.2	nc
Mussel Larvae	16%	58.6	84.4	nc
Mussel Larvae	18%	61.2	85.5	nc
Mussel Larvae	20%	63.6	86.5	nc
Sea Urchin Larvae PW	1%	89.9	94.4	89.3
Sea Urchin Larvae PW	2%	90.9	95.2	nc
Sea Urchin Larvae PW	3%	91.7	95.7	nc
Sea Urchin Larvae PW	4%	92.2	96.1	nc
Sea Urchin Larvae PW	5%	92.7	96.4	95.0
Sea Urchin Larvae PW	6%	93.2	96.7	nc
Sea Urchin Larvae PW	7%	93.7	96.9	nc
Sea Urchin Larvae PW	8%	93.9	97.1	nc
Sea Urchin Larvae PW	9%	94.2	97.3	nc
Sea Urchin Larvae PW	10%	94.3	97.5	97.6
Sea Urchin Larvae PW	12%	94.7	97.8	nc
Sea Urchin Larvae PW	14%	95.2	98.1	nc
Sea Urchin Larvae PW	16%	95.5	98.3	nc
Sea Urchin Larvae PW	18%	95.8	98.5	nc
Sea Urchin Larvae PW	20%	96.0	98.7	nc
Sea Urchin Larvae SWI	1%	79.4	81.7	62.4
Sea Urchin Larvae SWI	2%	81.4	83.5	nc
Sea Urchin Larvae SWI	3%	82.6	84.6	nc
Sea Urchin Larvae SWI	4%	83.5	85.5	nc
Sea Urchin Larvae SWI	5%	84.3	86.2	77.6
Sea Urchin Larvae SWI	6%	85.0	86.7	nc
Sea Urchin Larvae SWI	7%	85.5	87.3	nc
Sea Urchin Larvae SWI	8%	86.0	87.7	nc
Sea Urchin Larvae SWI	9%	86.4	88.1	nc
Sea Urchin Larvae SWI	10%	86.7	88.5	84.2

Table 15. Continued.

Test	p value	Tolerance Limits		
		Parametric		Non-Parametric
		Sources of Variation:		
		Multiple	Single	
Sea Urchin Larvae SWI	12%	87.4	89.2	nc
Sea Urchin Larvae SWI	14%	88.0	89.7	nc
Sea Urchin Larvae SWI	16%	88.6	90.3	nc
Sea Urchin Larvae SWI	18%	89.2	90.8	nc
Sea Urchin Larvae SWI	20%	89.6	91.2	nc
Neanthes Survival	1%	-5.9	42.2	-31.8
Neanthes Survival	2%	2.1	47.5	nc
Neanthes Survival	3%	9.1	50.8	nc
Neanthes Survival	4%	13.6	53.3	nc
Neanthes Survival	5%	17.2	55.4	18.2
Neanthes Survival	6%	19.0	57.1	nc
Neanthes Survival	7%	22.2	58.6	nc
Neanthes Survival	8%	23.3	59.9	nc
Neanthes Survival	9%	26.7	61.1	nc
Neanthes Survival	10%	27.5	62.2	39.7
Neanthes Survival	12%	31.8	64.2	nc
Neanthes Survival	14%	33.5	66.0	nc
Neanthes Survival	16%	36.7	67.5	nc
Neanthes Survival	18%	40.3	69.0	nc
Neanthes Survival	20%	42.4	70.3	nc
Neanthes Growth	1%	-1.5	39.6	-32.4
Neanthes Growth	2%	8.5	44.8	nc
Neanthes Growth	3%	13.6	48.2	nc
Neanthes Growth	4%	18.4	50.6	nc
Neanthes Growth	5%	22.8	52.7	17.3
Neanthes Growth	6%	25.4	54.4	nc
Neanthes Growth	7%	28.1	55.9	nc
Neanthes Growth	8%	30.6	57.2	nc
Neanthes Growth	9%	32.8	58.4	nc
Neanthes Growth	10%	34.4	59.5	38.8
Neanthes Growth	12%	37.6	61.4	nc
Neanthes Growth	14%	40.6	63.2	nc
Neanthes Growth	16%	43.3	64.7	nc
Neanthes Growth	18%	45.7	66.2	nc
Neanthes Growth	20%	47.9	67.5	nc
Nebalia	1%	-4.6	41.1	-31.2

Table 15. Continued.

Test	p value	Tolerance Limits		
		Parametric		Non-Parametric
		Sources of Variation:		
		Multiple	Single	
Nebalia	2%	5.5	46.2	nc
Nebalia	3%	12.0	49.3	nc
Nebalia	4%	16.5	51.7	nc
Nebalia	5%	19.5	53.6	16.2
Nebalia	6%	21.9	55.3	nc
Nebalia	7%	24.6	56.7	nc
Nebalia	8%	27.7	58.0	nc
Nebalia	9%	29.2	59.1	nc
Nebalia	10%	31.1	60.1	36.6
Nebalia	12%	35.1	62.0	nc
Nebalia	14%	37.7	63.7	nc
Nebalia	16%	40.9	65.1	nc
Nebalia	18%	43.1	66.5	nc
Nebalia	20%	45.7	67.7	nc

Table 16. Variance components: factors affecting variation in reference site toxicity data, and the percentage of variation accounted for by each factor. Values are computed from all data (this study, BPTCP screening studies, and SF Bay Regional Monitoring Program), with outliers excluded. Means and sd are given as a percentage of controls. Naive SD is the commonly used standard deviation, used to calculate tolerance limits with naive variance; SD is the standard deviation used in the bootstrap calculations, and is the square root of the sum of space, time, interaction, and replicate variance components (see Methods Section 10.3).

Protocol	Population Parameters			Variance Components			
	Mean	Naive SD	SD	Time	Space	Interaction	Error
Ampelisca	96	10.3	11.7	14%	0%	45%	41%
Eohaustorius (homog)	85	8.5	9.1	0%	10%	32%	58%
Eohaustorius (intact)	74	11.8	12.9	0%	45%	55%	0%
Eohaustorius (pw)	108	20.3	25.9	51%	31%	0%	18%
Mussel	101	11.4	15.4	36%	0%	47%	17%
Sea Urchin (pw)	102	2.5	2.8	53%	0%	16%	31%
Sea Urchin (swi)	98	4.7	4.8	37%	1%	62%	0%
Neanthes Survival	91	13.6	21.3	30%	25%	0%	45%
Neanthes Growth	88	13.5	20.2	13%	0%	65%	22%
Nebalia	88	9.3	16.6	38%	0%	62%	0%

Discussion

1.0 EVALUATION OF SEDIMENT REFERENCE SITES IN SAN FRANCISCO BAY

1.1 Sediment Chemistry

It is unlikely that there are any pristine sites in San Francisco Bay that would be indicative of pre-industrial conditions. All candidate reference sites evaluated in this study had detectable levels of numerous anthropogenic chemicals (Appendix A). All sites had nickel concentrations above PEL and ERM values (Probable Effects Level [MacDonald, 1994] and Effects Range Median [Long et al., 1995]), and all sites had chromium concentrations above PEL values. It is likely, however, that nickel and chromium were derived primarily from natural geologic sources, such as serpentine rock formations. Flegal et al. (1994) found that the concentrations of chromium and nickel in San Francisco Bay sediments were generally below their average crustal abundances, indicating they were not significantly enriched through human activities.

The polycyclic aromatic hydrocarbon (PAH) Dibenz[a,h]anthracene was measured at a concentration slightly above the PEL value (but below the ERM value) in one sample from Paradise Cove (Station 1, 3/95; Table 9). That sample did not elicit toxicity with any of the test protocols (Table 10). One sample from San Pablo Bay Island #1 had p'p'DDT and total DDT at concentrations well above both the ERM and PEL values (Table 9). Both of these samples were collected during the heavy storm events of March and April, 1995, and the elevated concentrations may have been associated with storm water runoff. However, in a replicate analysis of the Island #1 sample there was no detectable p'p'DDT or total DDT, though concentrations of other analytes were consistent with the original analysis (see Results Section 2.0). The distribution of DDT within this sample was apparently highly variable, and the original measurement may have detected a small amount DDT embedded within small sediment particles. Therefore, the toxicological significance of the measured DDT at Island #1 is uncertain.

The ERM values that were exceeded in some reference site samples from this study were among those for which Long et al. (1995) had limited confidence. ERM values for nickel, p,p'-DDT and total DDT were judged to have relatively low accuracy. Both nickel and the DDT compounds had low incidences of effects in studies where sediment concentrations were above the ERM. Chromium had high incidences of effects at concentrations above the ERM value, but this was exaggerated by data from multiple tests performed at only two sites (Long et al., 1995). Anderson et al. (1995) found that San Francisco Bay pore water had to be spiked with nickel to concentrations well above ERM values to elicit toxicity. The lack of significant toxicity in the

San Francisco Bay reference site samples with elevated nickel, chromium and/or DDT concentrations is, therefore, not without precedent. Because nickel was presumed to be derived primarily from natural sources, and because the measured nickel concentrations were below those expected to elicit toxicity (Anderson et al., 1995), nickel ERM quotients were excluded from the quotient means used to indicate the relative degree of pollution at the reference sites.

Since all sites had detectable levels of numerous chemicals, guideline quotient mean values were used as relative measures of the overall pollution at each site (see Methods Section 10.3.1). As above, guideline quotient values for nickel were not used in determining quotient means. The PEL quotient means for candidate reference sites ranged from 0.28 to 0.37. ERM quotient means ranged from 0.16 to 0.27. ERM quotients for the three RMP sites included in Reference Envelope calculations ranged from 0.09 to 0.11. The highest mean quotient values came from the San Pablo Bay Island #1 site, primarily as a result of high DDT measured in one of four samples from that site (Table 9). Use of guideline quotient means in the evaluation of sediment contamination has been limited, and interpretation is preliminary. In a recent study of San Diego Bay conducted as part of the Bay Protection and Toxic Cleanup Program (BPTCP), reference sites selected for use in statistical analyses of toxicity test data had ERM quotient means ranging from 0.065 to 0.252, and PEL quotient means ranging from 0.116 to 0.404 (Fairey et al., 1996). The ERM quotient means for all other samples analyzed in that study ranged from 0.088 to 2.373, and PEL quotient means ranged from 0.150 to 3.082. Four sites were identified in the San Diego study as having both toxic sediments and degraded benthic communities. For these four sites, the ERM quotient means averaged 1.47 (± 0.76) and PEL quotient means averaged 1.92 (± 1.01). All sites in that study having ERM quotient means greater than 0.55 had benthic communities that were classified as degraded (Fairey et al., 1996). The ERM and PEL quotient means from the five candidate reference sites in San Francisco Bay were low relative to quotient means from degraded sites, though some were higher than those from reference sites in the San Diego study. Excluding nickel and chromium, there were two San Francisco Bay reference site samples out of eighteen analyzed that had single chemical concentrations exceeding ERM and/or PEL values (Table 9), and numerous anthropogenic chemicals were detected at every site. The candidate San Francisco Bay reference sites, therefore, are clearly not pristine, but they may represent the best current characterization of optimal ambient conditions likely to be found in the Bay with reasonable sampling effort.

1.2 Salinity, Grain Size, and Total Organic Carbon

Salinity varied among sampling periods, especially in samples from San Pablo Bay, where pore water salinity ranged from 2‰ to 25‰ (Table 3). For all SF Bay reference site samples, the

salinity range (2 to 30‰) was fairly large, and was probably sufficient for comparisons with most test sites. Grain size distributions were similar to those at depositional sites suspected of having pollutant levels capable of producing biological impacts (e.g., Islais Creek, Castro Cove; Figures 16 to 23). All of the reference sites had high percentages of silt and clay. Because fine grained sediments are capable of scavenging and sequestering trace metals and other pollutants, it is important to have fine-grained reference sediments so that interpretation of observed differences in toxicity among sites can be attributed to factors other than grain size. TOC was relatively consistent among reference sites and test sites such as Castro Cove and Clipper Cove. Islais Creek sediments, which have received organically enriched effluents, had about twice the TOC content as reference site sediments.

The three RMP sites included in Reference Envelope calculations had generally larger grain size and lower TOC than did the reference sites evaluated in this report. Over two seasons (1996), percent fines (silt plus clay) and percent TOC, respectively, averaged 60% and 1.3% at Point Pinole, 36% and 0.8% at Horseshoe Bay, and 72% and 1.0% at San Bruno Shoal (SFEI, 1997).

1.3 Benthic Community Analyses

Sediment quality is commonly characterized using a triad approach that includes measures of chemistry, toxicity, and benthic community ecology (Chapman et al., 1987). Ideally, reference site sediments should be characterized using all three types of measurements. In San Francisco Bay, however, salinity fluctuations and invasions by exotic species have made it difficult to routinely characterize benthic communities (Nichols and Thompson, 1985), and this complicates efforts to make inferences about pollution impacts. Relationships between pollution levels and benthic community assemblages in San Francisco Bay have recently been the focus of SF Bay Regional Monitoring Program (RMP) pilot studies. Through cooperative efforts, these studies have included analyses of the reference sites at San Pablo Bay Island #1, Tubbs Island, and Paradise Cove. The results of these studies are presented in the RMP 1996 report (SFEI, 1997). That report states that firm conclusions about the condition of the benthos of the Estuary related to sediment chemistry cannot be made at this time, but the available data for the three reference sites indicate possible pollution effects. The San Pablo Bay Island #1 site had higher percentages of species characteristic of impacted sites, and lower percentages of species characteristic of unimpacted sites, than were observed in many other sites sampled during the RMP study. This suggests that the benthic community there may be affected by pollutants. The opposite indications were found at Paradise Cove, which apparently has a relatively unimpacted benthic community. The third reference site analyzed, Tubb's Island, was intermediate between the two in terms of both positive and negative ecological indicators. Of the three RMP sites included in

Reference Envelope calculations, Point Pinole appeared to have a moderately impacted benthic community, with indicator species distributions similar to Island #1. The benthic fauna of Horseshoe Bay and San Bruno Shoal appeared to be less impacted. These data, though preliminary, tend to characterize the reference sites as neither pristine nor severely impacted.

1.4 Toxicity Data

Toxicity test results have been used in previous studies to select and evaluate reference sites (PTI, 1991). However, without knowledge of the causes of sediment toxicity, it may be inappropriate to base reference site selection solely on the results of toxicity tests. Unexplained toxicity has been described in sites remote from sources of pollution (Long et al., 1990; PTI, 1991), and toxicity due to non-anthropogenic factors may be possible, though to our knowledge this has never been demonstrated. If reference sites are selected in advance of the sampling events in which they will be used for comparison with test sites, it seems reasonable for the selection process to include available toxicity data along with chemical, ecological, physical, and geographical information. Selection of reference sites based strictly on picking the least toxic sites from a single sampling event is difficult to justify.

In this study, nine toxicity test protocols were used in the evaluation of potential reference sites. Three of these protocols were included for specific study objectives, such as screening for TIEs or evaluating the effects of natural factors (*Eohaustorius* in porewater, *Eohaustorius* in intact cores, and *Nebalia*). As discussed below, the results from these three tests were variable and of limited use in reference site evaluation. Results from the other six protocols (*Ampelisca*, *Eohaustorius*, and *Neanthes* in homogenized sediment, mussel and sea urchin larvae in pore water, and sea urchin larvae at the sediment water interface) indicate generally high rates of survival, growth, or larval development at the reference sites (Figures 9 to 13). There were exceptions to this trend. Growth rates of the polychaete worms (*Neanthes*) were between 59% and 66% of the control value in samples from the two San Pablo Bay sites collected in Spring of 1995 (Figures 9 and 10, Table 10). The results of *Neanthes* growth tests did not correlate with any of the physical or chemical parameters measured, and the cause(s) for this response are not known. Survival of amphipods was depressed in individual field replicate samples, especially those from San Pablo Bay reference sites (Table 10). *Eohaustorius* data from all sites correlated with arsenic, copper, TOC, and percent clay. However, shifts in grain size and TOC at the San Pablo Bay sites did not appear to be related to the occasional observed decreases in survival (Figures 16 and 17; Appendix A), and there were no chemical analyses conducted on samples collected in the September 1994 survey, when most of the lower survival results were observed. For all San Francisco Bay reference site samples, larval development tests in pore water and at

the sediment water interface produced results similar to those in test controls. In general, the toxicity test results from reference site samples indicated no severe toxicity, slight to moderate toxicity in some samples from the two San Pablo Bay sites, and high rates of survival, growth and normal development at the remaining S.F. Bay sites. These trends are consistent with those from available chemical and ecological data, and indicate that some reference sites may exhibit moderate toxicity, but as a Bay-wide group they are probably representative of the least impacted conditions likely to be encountered in surveys of San Francisco Bay sediments.

Additional toxicity data from BPTCP screening surveys and RMP sampling were used in calculating sediment toxicity tolerance limits (see Table 14 and Discussion Section 3, below). Some of these data were identified as outliers, as will be discussed.

1.5 Tomales Bay and Bolinas Lagoon

Tomales Bay and Bolinas Lagoon were not used in calculations of reference envelope tolerance limits for San Francisco Bay. These sites may not be representative of San Francisco Bay sediment conditions (see for example Figure 21), and past and present data indicate the occurrence of unexplained toxicity at these sites (Flegal et al., 1994; Long et al., 1990; Table 10). As reference sites identified within San Francisco Bay appeared suitable for comparison with test sites there, the need for further investigation and use of remote reference sites diminished.

2.0 EVALUATION OF TOXICITY TESTS

2.1 Tests with the Amphipod *Eohaustorius estuarius*

Eohaustorius tests conducted according to the standard protocols (ASTM, 1993; USEPA, 1994) in homogenized solid-phase sediment met all criteria for toxicity test methods appropriate for use in San Francisco Bay as defined in this study. Control response was acceptable in all trials, the test was able to distinguish between sites with low and high concentrations of pollutants, there was low variability among laboratory replicates, and *Eohaustorius* is euryhaline. This amphipod has reasonable tolerance to ammonia and hydrogen sulfide (USEPA, 1994; Knezovich et al., 1995), and is more tolerant of fine grained sediment than the commonly tested amphipod *Rhepoxynius abronius* (USEPA, 1994). In this study, however, *Eohaustorius* survival correlated negatively with sediment clay/colloid content, especially in samples from Tomales Bay, and may be negatively affected by very fine grained sediment.

The *Eohaustorius* test in pore water had poor control survival and high variability. This test was not intended to be used to determine sediment toxicity at test sites, however, but rather to

investigate relationships between toxicity and chemistry through TIEs. The low control survival rates in some runs of this test were likely the result of stress to this infaunal organism caused by exposure to the water-only test conditions for extended periods (10 days).

The *Eohaustorius* test in intact cores had poor control survival and high variability. Intact cores were used in this study for comparison with homogenized samples to investigate the effects of sample homogenization. Homogenization of test sediments disrupts chemical equilibria and oxidation state, possibly causing artifacts that might influence test results. In this study, however, carnivorous annelids much larger than the test amphipods were occasionally observed in the intact core samples. These predators (and other organisms) are probably destroyed during the homogenization process, but predation may have had a significant effect on results of intact core tests. Elimination of interferences from other organisms in intact samples is the current obstacle to successful use of this exposure system for amphipods. Attempts have been made to eliminate interfering organisms through freezing and gamma irradiation of intact samples (Day et al., 1995), and other techniques such as use of microwaves or temporary elimination of dissolved oxygen may prove effective. Pursuit of these techniques would be worthwhile only if they were shown to be less disruptive than homogenization.

2.2 Tests with the Amphipod *Ampelisca abdita*

Control acceptability in tests with *Ampelisca* varied with organism source. While *Ampelisca* collected from San Francisco Bay have been tested successfully by other laboratories, tests using these organisms at MPSL resulted in control survival between 80 and 90%, less than the 90% criterion, despite previous experience with the protocol (> 5 sets of samples tested). Control survival of greater than 90% was achieved in both tests in which *Ampelisca* were obtained from the east coast. The *Ampelisca* test (ASTM, 1993; USEPA, 1994) distinguished sites having low and high concentrations of pollutants, demonstrated low variability among laboratory replicates, and was not affected by fine-grained sediments. This species is moderately tolerant of ammonia and hydrogen sulfide (USEPA, 1994; Knezovich et al., 1995). *Ampelisca* have been introduced to San Francisco Bay, and often occur in extremely high densities (Nichols and Thompson, 1985; SFEI, 1997), but *Ampelisca* is not as euryhaline as *Eohaustorius*.

2.3 Tests with the Mussel *Mytilus spp.*

The *Mytilus* larval development test in pore water met most test acceptability criteria. Four of five test series conducted in this study had acceptable control response. Unfortunately, the test series with unacceptable control response contained the Islais Creek and Castro Cove samples, so this evaluation of test sensitivity could not be made. This test is known to be sensitive to a

number of toxicants, however (e.g., Martin et al., 1981). *Mytilus* are more tolerant of estuarine salinities than are sea urchins, allowing their use in unadjusted pore water samples from a potentially larger portion of Bay sites, especially in site-specific studies. Salinity adjustment was often necessary in this study, however, because of the desire to test all samples from a given survey at the same salinity. Since salinity adjustment with brine involves different levels of sample dilution depending on original salinity, the results of porewater tests were often not directly comparable between sites. *Mytilus* is native to San Francisco Bay. The larvae are sensitive to hydrogen sulfide (Knezovich et al., 1995) and ammonia, which makes mussel test results difficult to interpret when these compounds are present at moderate concentrations.

2.4 Tests with the Sea Urchin *Strongylocentrotus purpuratus*

Sea urchin tests in pore water and at the sediment-water interface (SWI) had acceptable control response in all trials and low variability among laboratory replicates. While the two tests were strongly affected by samples from Islais Creek, hydrogen sulfide concentrations in those samples were sufficient to cause the observed result. However, TIEs demonstrated that this species was sensitive to Islais Creek samples even after ammonia and hydrogen sulfide concentrations were reduced to non-toxic levels (Hansen, 1996). Samples from Castro Cove were not toxic to the sea urchins. Tests with this species have been shown to be sensitive to a variety of toxicants (Bay et al., 1993). Many estuarine samples require salinity adjustment for pore water testing with *Strongylocentrotus,* which is a marine species. Since salinity adjustment with brine involves different levels of sample dilution depending on original salinity, the results of porewater tests were often not directly comparable between sites. Sea urchin larvae are sensitive to ammonia (Bay et al., 1993) and hydrogen sulfide (Knezovich et al., 1995). These chemicals can be accurately measured (Phillips et al., 1997) and removed from samples prior to testing, but selective removal is not always practical in survey studies, and other chemicals may be removed in the process. The sea urchin test did not appear to be affected by grain size, since larvae were not exposed to particles directly.

2.5 Sediment-Water Interface (SWI) Tests

Testing embryo/larval stages at the sediment-water interface offered some advantages over porewater testing. SWI tests were conducted on solid-phase samples with less than marine salinity because 33‰ overlying water was used. The salinity range for the SWI method has not been firmly established, but it probably could be used on samples from throughout the Bay. Salinity adjustment was uniform for all samples, because overlying water from the same source was used for all samples regardless of original sample salinity. This allowed direct comparability of results from different sites. Use of SWI exposure systems decreased the effects of ammonia

and hydrogen sulfide relative to porewater exposures (Table 12). This system may not be as protective of sensitive interstitial organisms, since chemical concentrations may be higher in sediment porewater than at the interface. It is likely, however, that SWI tests provide a more realistic sediment exposure for the developing echinoderm and bivalve embryos used in this study; they do not naturally occur in porewater, but are very likely to undergo some embryo development in contact with the sediment surface (Anderson et al., 1996).

2.6 Tests with the Polychaete Worm *Neanthes arenaceodentata*

The *Neanthes* test met control acceptability requirements in two of three sets of tests. *Neanthes* survival was significantly negatively correlated with the percentage of fine-grained sediment at the reference sites. While previous studies relating *Neanthes* growth and survival to sediment grain size have indicated a broad tolerance to sediments composed of 5 to 100% sand (Dillon et al., 1993), many of the reference sites had very fine grained material (Figures 16 to 23), perhaps beyond the range previously described. *Neanthes* has been shown to be tolerant of high ammonia and hydrogen sulfide concentrations (Dillon et al., 1993), and was not affected by sediments with high concentrations of these compounds in this study (e.g.. Islais Creek, Figure 26). The *Neanthes* test can be conducted at salinities as low as 20‰ (Dillon et al., 1993), allowing its use on sediments from throughout much of the estuary. Previous data suggest that the *Neanthes* test is less sensitive to a variety of toxicants and sediments than are the tests discussed above (Reish and Gerlinger, 1984; Anderson et al., 1998). In the present study, the *Neanthes* test made no distinctions between suspected polluted sites (Islais Creek and Castro Cove) and reference sites (Figure 26). High concentrations of ammonia and sulfide may have been responsible for Islais Creek sample toxicity in tests with some other species, but preliminary toxicity identification evaluations (TIEs) indicated that other chemicals were present in concentrations toxic at least to sea urchins (Hansen, 1996). The *Neanthes* test was apparently insensitive to ammonia, hydrogen sulfide, and other toxins in this sample. This relative insensitivity makes the test less useful in identifying sediments capable of producing biological effects, but could be advantageous at heavily polluted sites where high ammonia and sulfide concentrations preclude the use of other protocols.

2.7 Tests with the Leptostracan Crustacean *Nebalia pugetensis*

The Leptostracan crustacean *Nebalia pugetensis* was used in 10-day solid phase sediment tests because initial experiments indicated a high level of tolerance to hydrogen sulfide. Control survival failed to meet acceptability criteria in two of three tests, due primarily to poor organism condition at test initiation. It is likely that continuing effort could result in an adequate supply of acceptable test organisms, but this effort would have to be warranted by a demonstrated

advantage to using this test. The *Nebalia* test did not respond to the Castro Cove or Islais Creek sediments, validating its tolerance to ammonia and sulfide, but calling into question whether the test is sensitive enough to be useful in identifying problem sites. No further program effort is planned for test development with this species.

2.8 Correlations Between Sediment Chemistry and Test Results

Results of tests with *Eohaustorius* and sea urchins correlated significantly with contaminant concentrations at reference sites (Table 13). However, the biological significance of these correlations must be limited because of the generally low toxic response observed and the generally low concentrations of measured chemicals. Of the tests conducted, the *Eohaustorius* pore water test had the greatest variation in response, providing greatest resolution to allow statistically significant correlations. It is also possible that amphipod stress in water-only exposures increased sensitivity to the low concentrations of measured contaminants in correlation analyses.

Sea urchin and *Eohaustorius* test data also correlated with concentrations of total organic carbon (TOC). This result was similar to that observed in San Francisco Bay samples by Flegal et al. (1994). Many trace metals and non-polar organic compounds have an affinity for suspended and dissolved organic carbon, and TOC concentrations often covary with concentrations of a number of contaminants. Therefore, in large and diverse data sets, TOC often remains the last factor significantly correlated with toxicity, because it is present at all sites, while toxic covariants change from site to site.

The key to understanding relationships between sediment chemistry and toxicity is bioavailability. As in most sediment assessments, bulk sediments, rather than pore water, were analyzed chemically in this study, allowing greater uncertainty regarding partitioning and bioavailability of contaminants. Concentrations of acid volatile sulfide (AVS) were not measured, though AVS has been useful in interpreting relationships between toxicity and concentrations of some cationic metals in anaerobic sediments.